Building Student Ownership and Responsibility: Examining Student Outcomes from a Research-Practice Partnership

Marisa Cannata, Vanderbilt University
Christopher Redding, University of Florida
Tuan D. Nguyen, Kansas State University

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C100023. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

This is an Accepted Manuscript of an article published online by Taylor & Francis in Journal of Research on Educational Effectiveness on date, available at https://www.tandfonline.com/doi/full/10.1080/19345747.2019.1615157 on July 31, 2019.

Despite decades of ambitious high school reform, substantial evidence demonstrates reforms are inconsistently implemented and struggle to impact student learning (Datnow, Hubband, & Mehan, 2002; Mazzeo, Fleischman, Heppen, & Jahangir, 2016). In response, there has been a proliferation of new approaches to achieving school improvement at scale, such as improvement science and design-based implementation research (Bryk, Gomez, Grunow, & LeMahieu, 2015; Cohen-Vogel et al., 2015; Fishman, Penuel, Allen, & Cheng, 2013). While these methods differ in specifics, they share an assumption that improvement at scale comes not from replicating a proven program, but by practitioners and researchers working together with iterative, continuous improvement approaches to design and implementation (Bryk et al., 2015; Cohen-Vogel et al., 2015; Fishman et al., 2013).

These new approaches to scale reflect an increasing demand for a new type of research and design infrastructure like research-practice partnerships (RPPs) (Tseng & Nutley, 2014). RPPs are long-term, mutualistic, and intentionally structured collaborations between researchers and practitioners that bring original, rigorous research to bear on a particular problem of practice

(Coburn, Penuel, & Geil, 2013). RPPs can take multiple forms, such as research alliances, design-research projects, and Networked Improvement Communities (NICs), the latter of which has particularly grown in prominence due to work by the Carnegie Foundation for the Advancement of Teaching to advance improvement science and continuous improvement approaches (Coburn et al., 2013; Cohen-Vogel et al., 2015). With an increasing array of funding sources emphasizing RPPs, such as the Institute of Education Sciences (IES), National Science Foundation, Spencer Foundation, and William T. Grant Foundation, research on RPPs is accumulating. For example, innovations developed through research-practice partnerships have been tested in rigorous efficacy studies with desirable outcomes (Booth et al., 2015; Sowers & Yamada, 2015). There is also research on the challenges of engaging in RPPs and the internal structures that support RPP work (Coburn & Penuel, 2016; Conaway, Keesler, & Schwartz, 2015; López Turley & Stevens, 2015), including the role of rapid-cycle continuous improvement processes in NICs (Hannan, Russell, Takahashi, & Park, 2015; Russell et al., 2017; Tichnor-Wagner, Wachen, Cannata, & Cohen-Vogel, 2017).

While the research base on RPPs in general, and NICs in particular, is growing, so is the recognition that improving these partnerships require greater attention to how specific dynamics of RPPs are related to the outcomes they achieve (Coburn & Penuel, 2016). That is, in the words of IES director Mark Schneider (2018), research is needed "[identifying] the functions, structures, or processes that work best for increasing the *impact* of RPPs." In this paper, we report evidence of student outcomes from a multi-year partnership within one large, urban district. In this partnership, we established a NIC with a shared theory of improvement and where three schools co-developed practices to improve student ownership and responsibility using a continuous improvement process. We evaluate evidence by assessing changes in grades,

course failures, discipline, and attendance. We adopt a mixed methods framework to describe both evidence of student outcomes and the features of the RPP and improvement approach that may have shaped these outcomes. We seek to answer two research questions:

(1) To what extent did the co-developed practices reduce students' disciplinary infractions and the number of failed courses and improve student grades and attendance?

(2) How do specific features of our improvement approach (shared theory of improvement, rapid-cycle testing, and research-practice partnership) explain differences in implementation quality and observed outcomes?

We begin by describing the three core features of our improvement approach and situate those features within the broader literature on NICs and RPPs. We then detail the specific context of our partnership and how these improvement features were enacted. Next, we describe the data used for this study as well as the quantitative and qualitative methods. We then present our results, first providing quantitative evidence on four student outcomes: attendance, discipline, grades, and course passing. We then describe how each school enacted the improvement approach and their level of implementation to explain school-level differences in outcomes.

Networked Improvement Communities and Continuous Improvement

Several models of RPPs have been proposed to address the contextual factors that shape implementation and scale up (Coburn, Penuel, & Geil, 2013). Understanding how the structures and processes of RPPs contribute to student outcomes requires unpacking the different features of RPPs and theorizing how those features contribute to positive outcomes. As our partnership was structured as a Networked Improvement Community, we draw heavily from four frameworks about RPPs and NICs (Coburn et al., 2013; Cohen-Vogel, Cannata, Rutledge, & Socol, 2016; Henrick, Cobb, Penuel, Jackson, & Clark, 2017; Russell et al., 2017). Looking

across these frameworks and the literature on scaling up school reform, we identify three core features of NICs that seem particularly poised to shape the successful implementation and scaling of improvement initiatives: deep understanding of a theory of improvement, rapid-cycle testing, and building educator capacity to engage in the network and lead improvement in their school. Further, we argue that the ways in which NIC work shapes outcomes depends on both partnership-level activities and school-level activities. Frameworks to assess RPPs attend to both indicators that reflect the partnership overall, such as communication processes and research infrastructure, and the work of individual members, such as capacity to engage in new roles. Figure 1 presents a visual representation of our framework and illustrates how both school-level and partnership-level features shape implementation and outcomes. We focus first on the school-level improvement features and how they contribute to successful implementation at scale.

Understanding the Theory of Improvement

RPPs are defined, in part, by a shared focus on a particular problem of practice (Coburn et al., 2013), and NICs are distinguished by their use of a theory of improvement. Russell and colleagues describe the importance of this shared theory of improvement by noting the "theory grounds the collaborative work of the NIC by specifying the problem and aim that the NIC is pursuing and unpacking the systemic context that produces the problem" (Russell et al., 2017, p. 17). Importantly, the theory of improvement is not devoid of context, but is grounded in the context in which the improvement work is occurring (Cohen-Vogel, Cannata, Rutledge, & Socol, 2016). A NIC's theory of improvement reflects both expertise from the research community and an understanding of the system that is producing the current problem (Henrick, Cobb, Penuel, Jackson, & Clark, 2017). The development of the theory of improvement that connects the ultimate goal with a shared understanding of the drivers that contribute to that goal is critical to

successfully launching a NIC (Russell et al., 2017).

Shared understanding of the theory of improvement is important for achieving success at scale because many school reform efforts result in superficial changes in classroom practices or grafting new practices onto old routines without shifts in deeper pedagogical principles (Elmore, 1996; Spillane, Reiser, & Reimer, 2002; Supovitz, 2008). Deep instructional change requires altering teachers' beliefs about how students learn, expectations for students, and the norms of interaction in schools and classrooms (Coburn, 2003). Educators are often not aware of how the theories of learning embedded in reform initiatives may conflict with their own unstated theories of learning, which then creates challenges for implementation (Hatch, 2002). On the other hand, when educators have a deep, internalized understanding of the ideas embedded in a reform, they can apply them in situations where the reform itself does not offer explicit guidance (Honig, Venkateswaran, McNeil, & Twitchell, 2014). Attending to how educators understand the theory of improvement underlying the reform initiative is even more important as NICs shift from a focus on fidelity of implementation to maintaining integrity with adaptive integration (Cannata & Rutledge, 2017; Hannan et al., 2015). Successful adaptations hinge on whether educators understand not only the innovation practices themselves, but the theory behind them (Dede, 2006; Thompson & Wiliam, 2008).

Rapid-cycle Testing

NICs seek to engage educators in collecting data for rapid-cycle improvement efforts and build up to larger scale change through continuous improvement (Coburn et al., 2013; LeMahieu, Grunow, Baker, Nordstrum, & Gomez, 2017; Russell et al., 2017). Bringing educators and researchers together to collaboratively design, study, and iterate on effective practices as educators adapt them to their specific contexts is a common element of a variety of emerging

approaches to collaborative reform initiatives (Cohen-Vogel et al., 2015). The Plan, Do, Study, Act (PDSA) cycle is one common approach, and requires identifying the aim of a particular improvement, testing the change idea, and monitoring whether the observed changes led to the intended improvement (Langley et al., 2009). Rapid-cycle testing should be iterative, as results from an individual test can lead to either revising and testing the change again, or deciding to scale it into more diverse contexts. Rapid-cycle testing should also be problem-focused and tied to the theory of improvement (Langley, 2009).

Rapid-cycle testing is an important component of this improvement paradigm because, while there are many innovations that have positive outcomes in rigorous efficacy trials, it is less clear whether these innovations are always usable for schools (Coburn & Penuel, 2016). Reforms that are not consistent with the local organizational context—no matter how effective they may be in controlled trials—face serious difficulties with implementation (Bodilly, 1998; Bryk, 2015; Elmore, 1996; Fullan, 2001). Further, educational implementation research has long noted that schools adapt innovations to focus on their unique needs, sometimes to ill effect (Datnow & Park, 2009; Siskin, 2016). Continuous improvement approaches to scale can bring discipline to the adaptation process as school teams share evidence of what they have accomplished with others focused on the same problem (Cannata, Cohen-Vogel, & Sorum, 2017; LeMahieu et al., 2017). Rapid-cycle testing can also address another challenge to scaling up reform—buy-in and ownership—because local practitioners are involved in developing and testing the innovation (Cohen-Vogel et al., 2016; Datnow, Hubbard, & Mehan, 2002). This attention to local context is particularly important for achieving scale as innovations must be able to fit with contexts that vary greatly while coping with change, promoting ownership, building capacity, and enabling effective decision-making (Cohen et al., 2013). At the same time, NIC members must engage in

rapid-cycle testing and adaptive integration through a disciplined process and have sufficient capacity (Russell et al., 2017). This disciplined approach to improvement ensures that educators are making evidence-based decisions about which practices to implement, and how to implement them, as multiple forms of evidence are examined to test and refine the practices in their context.

Educator Capacity to Engage in Partnership

Research on school reform demonstrates that successful implementation of change initiatives requires some existing capacity at the school-level (Hatch, 2002). Scholarship on RPPs makes clear that engaging in an RPP requires new roles for both researchers and practitioners. For example, Coburn, Penuel, and Geil describe a core component of NICs as having a "primary focus on developing local capacity" (2013, p. 13). Indicators of this dimension include whether members "develop professional identities" consistent with their work, "assume new roles and develop the capacity to conduct partnership activities," and experience "change in the practice organization's norms, culture, and routines around the use of research and evidence" (Henrick et al., 2017, p. 25).

The delineation of these different dimensions of capacity reflect the need for human, social, and cultural capital to engage in a NIC (Rubin, Nguyen, & Cannata, 2015). At both an individual and organizational level, schools need to have sufficient human capital, including the knowledge, skills, resources, and personnel to engage in the work expected of them (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Durlak & Dupre, 2008; Spillane et al., 2002). At the organizational level, a productive social infrastructure, such as a history of collaboration in the school, stability of faculty, and trust allow individuals to access social capital to support implementation (Bryk et al., 2010; Murphy & Torre, 2014; Redding, Cannata, & Taylor Haynes, 2017). Finally, capacity in a NIC involves a normative components as educators must adopt new

professional identities, suggesting that new forms of cultural capital need to be established (Henrick et al., 2017; Russell et al., 2017).

Partnership-level Improvement Features

School improvement research suggests understanding how these improvement features are enacted at the school-level help us predict the quality of implementation and ultimately student outcomes. Yet, within the context of a collaborative reform model, school-level features are supported by partnership-level dynamics. For example, building the theory of improvement based on evidence from participating sites is a core element of what the NIC's hub organization should be doing when launching the NIC (Russell et al., 2017). The other core elements are helping NIC members learn how to use improvement research methods and building the analytic infrastructure (Russell et al., 2017), both of which support school-level engagement in rapid-cycle testing. In addition to supporting the school-level improvement features, the partnership-level should also be focused on leading and operating the network as a whole, including establishing collective norms, maintaining the partnership's focus on learning and providing evidence that connects classroom to district-level processes (Coburn et al., 2013; Henrick et al., 2017; Jaquith, 2017; Russell et al., 2017).

The presence of both school- and partnership-level contributions to NICs brings methodological challenges when assessing the impact of the innovations developed through them. First, as all schools benefit (or suffer) from the partnership-level work, there is no way to distinguish the effect of this overall partnership from the specific practices that are developed. In other words, the treatment is comprised of both the partnership itself and innovation design. That said, the school-level improvement features can be assessed separately at each school. Second, while the innovation was rooted in a common theory of improvement, the use of rapid-cycle

testing and adaptive integration resulted in differences in how the innovation was implemented across these schools. Consequently, to provide lower and upper bounds of the effects of the innovation on course failure, grades, days absent, and number of disciplinary infractions, we adopt two estimation strategies—a gains model and a difference-in-differences estimation strategy to compare student outcomes among the innovation schools to the remaining schools in the district (Angrist & Pischke, 2009). To address the concern with school-level variation, we examine both the overall effect of participating in this collaborative reform initiative and differences among the innovation schools to examine how each school's enactment of the improvement features are associated with student outcomes. In the next section, we describe the particular context of the NIC from which the data are drawn.

Research-Practice Partnership: Developing the SOAR Innovation

This partnership began with intensive study of higher and lower performing high schools in the district, and identified Student Ownership and Responsibility (SOAR) as the key differentiating feature. The NIC was launched in 2012-13, when the SOAR design team was established. The SOAR design team was comprised of about 25 individuals, including teachers, school administrators, central office administrators, program developers, and researchers. The first author, whose background is in studying school reform and educational policy and led the research that identified SOAR, was part of the research team and on the district design team. The central office administrators included the deputy superintendent who oversaw teaching and learning, leadership, and student support; the data and accountability director; curricular specialists; and advanced academics. School-level members included assistant principals and teachers across subject areas. A retired principal served as a coordinator who supported logistics and acted as a liaison to both the research team and schools. The deputy superintendent and

coordinator recommended central office members of the design team. School-level members were recommended by their principal. In 2012-13, the design team met for two days every month to examine the initial research, conduct needs analysis related to SOAR, engage in capacity-building activities, and design a SOAR prototype aimed at creating norms and school-wide practices that foster learning and engagement among students (see Table 1 for timeline of design activities and meetings). These meetings were organized around two connected learning goals: learning about implementation and scale and learning about SOAR and its enactment within the district context. Thus, the district design team spent time building the theory of improvement around SOAR while also building capacity and norms to engage in the NIC.

The theory of improvement around SOAR was grounded in both the specific findings from this district and the broader literature on the importance of co-cognitive student attributes such as efficacy, problem-solving, and academic and behavioral engagement (Dweck, 2007; Fredricks, Blumenfeld, & Paris, 2004; Schunk, 1991). This focus on changing students' mindsets and providing them problem-solving skills to engage in academic work builds on a robust empirical research base on co-cognitive factors (Farrington et al., 2012). Specifically, SOAR focused on building a student growth mindset and developing problem-solving skills to improve student engagement.

With the districtwide team outlining the theory of improvement, each innovation school established a SOAR team in 2013-2014 to pilot practices and further develop them within their context. SOAR teams had 6-8 members, who were almost all teachers (one school had an assistant principal on the team). Each school team had a teacher in the Advancement Via Individual Determination (AVID) program, which was an existing district program considered to be related to SOAR. The SOAR team was responsible for leading implementation in their school,

often by working with the administration, developing SOAR practices, using PDSA, and providing training for other teachers in the school to enact SOAR practices. During this year, school teams met as a whole group once a month to deepen their knowledge of SOAR, learn how to engage in rapid-cycle testing, and share what they are learning through their PDSA cycles. The district design team continued to provide overall leadership for the network. Specifically, they organized trainings around PDSA, determined the capacities the network members needed, designed learning activities around those areas, and facilitated network sharing of what each school was learning. This contributed to revising the shared practices to develop SOAR in students. Through this development process, school teams were also given leeway in customizing the common district design to their particular context. While each school design team implemented these common elements of the design, its delivery varied in ways that may shape student outcomes. By the end of the 2013-14 development year, the core practices of the innovation included 1) teaching about growth mindset, 2) student grade monitoring and goal-setting activities, 3) problem-solving activities that supported students in improving their grades, and 4) a behavioral reflection form designed to get students to reframe problematic behaviors before creating a disciplinary referral. The final SOAR component focused on building a school culture around SOAR. Full implementation began in 2014-15 and continued in 2015-16.

The research team engaged in other activities that supported the NIC's work. One, the research team conducted visits to each innovation school to learn about implementation, and shared memos about implementation with the schools and district design team. More details on these visits are provided below. Two, the research team, along with the program developers, provided training and coaching to support teams in conducting rapid-cycle testing. Finally, the research team worked with the district design team to develop outcome indicators by which the

NIC would assess their overall progress. These outcome measures were designed to capture both shorter and longer-term outcomes that reflected the SOAR theory of improvement and were tied to important district goals. The long-term outcomes of GPA and course failure reflect the focus on co-cognitive traits. Recent research has indicated that high school course grades are better predictors of college access, college graduation, and longer-term life outcomes than test scores. GPA, for example, is a consistent predictor of graduating from both high school and college, and a "primary driver of differences by race/ethnicity and gender in educational attainment" (Farrington et al., 2012, p. 3). Further, failing a course predicts dropping out of high school (Bowers, Sprott, & Taff, 2013). The short-term measures are attendance and disciplinary infractions, which reflect the SOAR theory of action of academic and behavioral engagement. Improving attendance can also improve high school graduation and college enrollment (Faria et al., 2017; Mac Iver & Messel, 2013). In the district in this study, student disciplinary infractions cover a range of behaviors, such as bullying, fighting, or disrespect to teacher, but is often met with a similar outcome: in-school or out-of-school suspension. Such disciplinary action is associated with student grades and achievement test scores (Arcia, 2006).

Research Design

Study Sample

This southwestern district served approximately 80,000 students; the majority were low-income or from traditionally underserved racial/ethnic groups. The innovation schools were selected through a collaborative process with district personnel and school administrators. The selected schools expressed an interest and willingness to participate in this innovative reform model. While a school's value-added performance was not used in the selection of these schools, their school value-added suggests that they were moderately performing schools in the district.

Table 2 presents the characteristics of the three innovation schools and other high schools in the districts. Fewer innovation school students received free or reduced price lunch or identified as Black, although more innovation school students identified as Hispanic. Students from Hancock failed more classes and had lower average grades than students in non-innovation schools in the district. Students at Williams and Smith had higher grades than students in non-innovation schools and students at Smith also failed fewer courses. Students at Smith and Hancock also were absent less frequently. Compared to the district, Smith and Hancock had fewer Black students but more Hispanic students.

We use qualitative and quantitative data to understand outcomes of the partnership in one district over two years of implementation. The qualitative data for this study come from two sources: observations of NIC meetings and field visits in these three innovation high schools. During the 2014-15 and 2015-16 school years, we observed all 13 meetings where SOAR teams met together as an entire network. The first author, as a member of the design team, was a participant observer in these meetings, and the second and third authors (along with additional researchers) took fieldnotes, collected feedback forms from NIC members, and collected artifacts of documents shared or created during the meeting. Further, two four-day field visits occurred in the first year of implementation (October 2014 and April 2015) and one three-day visit in March 2016, the second year of implementation. Over these three visits, we conducted 9 principal interviews, 17 interviews with other administrators, 72 interviews with members of the SOAR team, 173 interviews with other teachers, 19 focus groups with teachers or support staff, and 34 student focus groups. We use the fieldwork data to provide evidence on enactment of the practices and how participants described the outcomes they were achieving as a result of this work. The interviews and focus groups focused on their understanding of student ownership and

responsibility and specific innovation practices, support for the innovation, the extent to which they enact SOAR practices, how the SOAR team worked as a group, the capacities of the SOAR team, and how they engaged in rapid-cycle testing. Interview and focus group guides are in the Appendix.

We also take advantage of rich administrative data from the district for all high school students enrolled in the 2010-2011 to 2015-2016 school years. The data used for this study includes 91,410 student-year observations. About 3% are dropped from the analysis due to missing data. The analytic sample includes 33,215 unique student observations.

Improvement Approach and Implementation Measures

Following each research visit, data were coded using an a priori framework for implementation that focused on facilitating conditions (will, capacity, beliefs about SOAR, and alignment to context), implementation supports (implementation team dynamics, engagement in rapid-cycle testing, leadership, resources/training), implementation quality which itself involved teacher experiences with implementation (enactment of innovation practices, feedback on practices) and student experiences with implementation (responsiveness, perceived outcomes). The coding team first coded several transcripts independently, and then compared coded transcripts to ensure they were applying codes consistently. Through multiple rounds, the coding framework was revised or clarified. For example, capacity was expanded to differentiate between capacity of teachers to enact SOAR practices, capacity of the implementation team to lead the work, and organizational capacity of the school.

Once the coding team agreed on the final coding scheme, they independently coded all transcripts. After coding was complete, a researcher prepared detailed memos for each school for each major theme in the coding framework. This process was repeated after each field visit. In

Year 2, the coding scheme was further expanded to include antecedents to sustaining and scaling the practices. Memos around these themes at each time point served as the primary documents for investigating the enactment of the improvement approach and quality of implementation. Specifically, three coders independently categorized each school on the three improvement features (understanding of the theory of improvement, engagement in rapid-cycle testing, and capacity to engage in the partnership). Detailed rubrics that guided these categories are included in Table A1. For understanding the theory of improvement, we sought out evidence that both the SOAR team and other school stakeholders demonstrated an understanding of SOAR and how the specific practices were theorized to contribute to student ownership. For rapid-cycle testing, we sought out evidence that the SOAR team's enactment of PDSA was problem-centered, iterative, used multiple forms of evidence, and resulted in evidence-based decisions on how to improve SOAR practices. For capacity to engage in the partnership, we sought out evidence that the SOAR team had the human, social, and cultural capital necessary.

For implementation quality, there were five SOAR practices for which we analyzed the quality of how the schools enacted the innovation practices: teaching growth mindset, goal-setting and grade monitoring practices, problem-solving practices, rewarding positive behavior, and building a school culture around these practices. For each of these practices, four researchers independently read memos on implementation quality and categorized each school as high, medium, or low enactment. High implementation quality existed when the practice was consistently implemented throughout the year. Medium implementation quality existed when the practice was implemented, but was inconsistent throughout the year. For example, a practice may have had high implementation in the beginning of the year, but waned over time. Low implementation quality reflected little to no indication this practice was implemented. For both

the improvement approach and implementation quality measures, researchers met to reconcile their independent categorization, using a consensus process to determine the final rating.

Outcome Measures

Outcomes include students' grades, passing rates, absences, and disciplinary infractions. Student's grades are the averages of the students' scores for each class. In 2013-2014, this measure ranged from 0 and 100, with an average student grade of 82 (see Table 2). When operationalizing a students' passing rate, we focus on the number of courses a student did not pass throughout the school year. Students were considered to be failing a course if they did not score at least a 70% in a course. As students could be registered for up to nine courses a semester, the maximum value for this variable is 18. Although the modal value for this variable is 0, on average, students did not pass 1 course. The measure for days absent is the number of days a student did not attend in a particular year. Student infractions is a measure of the number of infractions a student received in a particular school year. Infractions include code of conduct violations for behaviors such as cheating, disrespect towards teachers, bullying, fighting, disobeying school rules, dress code violations, or possession of tobacco. Infractions also include more serious offenses such as drug or alcohol use, criminal mischief, assault, arson, felony, possession of a weapon, public lewdness, gang violence, or serious misbehavior.

We also include controls for binary indicators of student race/ethnicity (Black, Hispanic, or other race/ethnicity), free and reduced lunch (FRPL) status, gifted status, and grade level. Additionally, we control for the number of days in which a student was enrolled in a school, indicators of whether or not they withdrew or started after the beginning of the school year, and for the number of courses in which a student is registered throughout the school year. At the school level, we control for student enrollment as well the proportion of Black students,

percentage of Hispanic students, and students who receive FRPL.

Methods

For this study, we adopt a sequential mixed methods research design (Smith, Cannata, & Taylor Haynes, 2016; Teddlie & Tashakkori, 2006). We first conduct the quantitative analysis to ascertain the extent to which students in the innovation schools benefited from the partnership. We then draw on qualitative fieldwork data to determine the degree of engagement in the NIC, quality of implementation, and participant understandings of accomplishments. This analytic process used several strategies to address potential threats to the validity of our inferences from the qualitative data, including cross-validation between researchers, triangulation among sources and perspectives, and member checking (Miles & Huberman, 1994; Patton, 2002). For example, we sought out comparisons between perspectives of the SOAR team and perspectives of others in the school, recognizing that overreliance on the SOAR team may reflect elite bias (Miles & Huberman, 1994). We also shared versions of the researcher-developed memos on implementation with the SOAR teams and district design team. Triangulating between the qualitative and quantitative findings also encouraged us to consider rival hypotheses.

For the quantitative analysis, we use a lagged dependent variable and difference-in-difference (DD) approach to give us plausible bounds on the estimated treatment effect of the SOAR Innovation under different assumptions (Angrist and Pischke, 2009). This ordinary least squares (OLS) gains model can be estimated:

$$Y_{ist} = \beta_0 + \beta_1 Innovation_{st} + \beta_2 Y_{ist-1} + \beta_3 \boldsymbol{X_{ist}} + \beta_4 \boldsymbol{S_{st}} + \gamma_t + \epsilon_{ist} \quad (1)$$

where Y_{ist} is the outcome for student *i* in school *s* in year *t* and $Innovation_{st}$ is a dummy

variable for whether and when the school implemented the SOAR innovation[1], Y_{ist-1} is the lagged dependent variable, $\boldsymbol{X_{ist}}$ is a vector of student controls, $\boldsymbol{S_{st}}$ is a vector of time-varying school characteristics, γ_t is a year fixed effect, and ϵ_{ist} is an error term. In this model, β_1 can be interpreted as gains in each outcome among students in the innovation schools in the post-treatment period.

This gains model would be biased by unobserved school-level factors that differ between innovation and non-innovation schools in the district and influence student performance on any of the outcome variables. To address this concern, we add to the model a school fixed effect (δ_s) to compare students' prior outcomes to average student outcomes in non-innovation schools. When innovation and non-innovation schools have a similar pre-treatment trend, the DD model represents the counterfactual change of implementing the co-developed innovation. This ordinary least squares (OLS) model can be estimated:

$$y_{ist} = \beta_0 + \beta_1 Innovation_{st} + \beta_2 \boldsymbol{X_{ist}} + \beta_3 \boldsymbol{S_{st}} + \delta_s + \gamma_t + \epsilon_{ist} \quad (2)$$

In this model, β_1 can be interpreted as the difference in student outcomes between innovation and non-innovation schools after implementation. To account for repeated observations of student over time, standard errors are clustered at the school level in Models 1 and 2 (Bertrand, Duflo, & Mullainathan, 2004).

This initial analysis estimates an average treatment effect for the students in schools that participated in this continuous improvement process. In addition to this overall treatment effect, we examine several heterogeneous treatment effects. These include differences across the three

[1] The post-treatment period does not include the year when the SOAR team developed and piloted the practices of the SOAR innovation. This decision is justified for two reasons. First, the piloting that did occur was limited to members of the SOAR team. Second, when a practice was piloted, it tended to only be implemented once or twice, limiting its potential impact on student outcomes. Nevertheless, it is possible that this piloting would lead to pre-treatment differences.

innovation schools and post-treatment year.

In additional analysis, we examine the robustness of the DD research design. An assumption of this research design is that innovation and non-innovation schools had similar pre-treatment trends in the outcome variables. We test for this assumption in two ways. We first estimate the relationship between innovation school participation and student outcomes not only in the post-treatment period but also in all years in which we have data. These estimates function as a placebo test where, conditional on covariates, any significant differences in the slope between innovation and non-innovation schools prior to treatment indicates a violation of the parallel trends assumption. Graphically, we also show the predictions from this regression to visually examine the presence of pre-treatment trends, when holding all other variables at their mean. Evidence of a violation of this assumption would indicate pre-treatment differences between innovation and non-innovation schools that could explain why innovation schools have more positive outcomes in the post-treatment period, outside of their participation in the improvement process. In general, we find evidence of parallel trends in terms of student grades and number of failed courses but not attendance and disciplinary infractions. In addition, we find evidence that Williams High School and Smith High School pre-treatment trends generally resembled the non-innovation schools. The evidence of pre-treatment differences at Hancock High School limit our inference of the effect of SOAR at this school. We discuss the full results of this sensitivity analysis below.

Findings

Impact of the Co-Developed Innovation on Student Outcomes

Our first research question asks about the extent to which the co-developed innovation reduced students' disciplinary infractions and the number of failed courses and improved student

grades and attendance. We find no evidence of an overall relationship between the SOAR innovation and student outcomes that is robust to model specification. However, when results are separated by school, Williams and Smith each saw increased student grades and fewer absences persisting across both years of implementation.

Table 3 reports the gains model and DD estimates of the four outcomes: days absent, the number of disciplinary infractions, the number of classes failed, and average grades (full results are in Appendix Table A2). In the gains model, the coefficient on innovation school indicates that students in these schools had average decreases in the number of days absent (-1.17, p = 0.02) and increases in their grades (0.95, p = 0.08). These slight improvements translate to relatively small effect sizes: a 0.04 decrease in days absent and a 0.04 standard deviation increase in average grades. At the same time, we find a slight *increase* in the average number of infractions (0.08, p = 0.002). However, when conditioning on unobserved school-level characteristics in the DD model, we find no evidence of a relationship between the SOAR innovation and any student outcomes.[2] This finding suggests any overall gains experienced by students in the innovation schools can be explained by unobserved but fixed school characteristics between innovation and non-innovation schools.

When the results are separated by innovation school (Table 4), we find notable heterogeneity across the innovation schools, with fairly robust evidence of improvements in Williams and Smith High Schools. In Williams High School, we find students were absent between 1.05 and 1.25 fewer days. Students' grades improved by 0.74 to 1.42 points, on average, depending on the model. We also show a positive effect on student grades, number of failed

[2] In Tables A3 and A4, we include the lagged dependent variable when estimating the school fixed effects model. In the Table A3 model, the estimates on the number of failed classes and average grades are significant. In Table A4, the estimates on the number of failed classes, average grades, and days absent are significant or marginally significant for all schools, with the exception of days absent for Smith High School.

classes, and absences in Smith High School. Students' grades improved by an average of 1.24 to 1.64 points, they failed between 0.33 and 0.44 fewer courses, and they were absent between 1.10 and 1.36 fewer days. For both schools, the estimates are less consistent for the number of disciplinary infractions in terms of the magnitude, direction, and level of significance. The estimated effects of the SOAR innovation are less consistent in Hancock High School. In the gains model, there is no evidence of a relationship between the SOAR innovation and grades or the course failure, a marginally significant decrease in absences, and a slight increase in disciplinary infractions. Estimates from the DD model show SOAR was linked with worse student outcomes, thereby offsetting the positive educational effects of the innovation in the other two schools.

In Appendix Tables A5 and A6, we examine whether these differences are driven by the first or second year of implementation, but generally find that the year 1 and year 2 effects are comparable, although not statistically significant in most cases, with the exception of the gains model predicting student absences. In Williams, increases in average grades and decreases in student absences were observed in both years and robust to both estimation strategies. In Smith, we find decreases in the number of failed classes and student absences and increases in grades in both years. At Hancock, the estimated treatment effect is fairly consistent across the two years but again not robust across the different estimation strategies.

Sensitivity Analysis

A concern with this analysis is that positive outcomes attributed to the innovation designed through the continuous improvement reform model are a result of the innovation schools being selected to participate in this program based on unobserved, time-varying characteristics. For example, schools selected to participate in this improvement process may

have unobserved, time-varying characteristics that would make them more likely to improve student outcomes, regardless of their participation in this process. If this scenario were true, we would be worried that the factors that led district stakeholders to select the innovation schools in the first place explain the school improvements rather than actual participation in the improvement process and the implementation of the SOAR innovation. A related concern is that innovation schools could have been selected based on past student outcomes. To the extent to which prior student outcomes were related to any transitory shock, any post-treatment improvements may arise from regression to the mean.

In Figure 2, we provided graphical evidence of the parallel trend assumption, comparing the predicted outcomes across all periods while holding all student and school characteristics at their mean. The results for failed classes, average student grades, and student absences suggest similar pre-treatment trends between non-innovation schools in the district and Williams and Smith, and for absences in Hancock. Hancock had lower failure rates that increased before and after the implementation of the SOAR innovation. The innovation schools did not consistently follow the pre-treatment trends for non-innovation schools in the district in terms of the number of disciplinary infractions.

Tables A7 and A8 further examine differences in pre-treatment trends. We find marginally significant evidence that innovation schools had fewer disciplinary infractions in 2014. When separated out by school, we find some evidence of pre-treatment differences in the trends of the outcomes, although the pre-treatment differences are most concentrated in the number of days absent and disciplinary infractions or at Hancock High School. Most notably, Hancock consistently had higher course failure rates, lower grades, a higher absentee rate, and a higher and lower disciplinary infraction rate, depending on the year. This suggests that Hancock

differed from the district in ways that could have shaped its uptake of the innovation and the resulting effect on students. We find similar evidence at Smith, although it is strongest for days absent and disciplinary infractions. As a result, we cannot rule out the possibility that the decreases in student absences following the implementation of SOAR in Smith were due to previous efforts to improve student attendance.

Looking across the gains and DD models, and considering the sensitivity analyses, we do not see any consistent evidence of an overall association between SOAR implementation and student outcomes. We do see consistent evidence that SOAR implementation was associated with decreased student absenteeism and improved student grades at Williams and Smith, as well as decreased course failure at Williams. That being said, at Smith, it is possible that improvements in student attendance are attributable, in part, to pre-treatment trends. There is no consistent evidence that is robust to sensitivity analyses that SOAR implementation was associated with any outcome in Hancock. We now turn to school-level engagement in the NIC to further explore how the improvement approach may have influenced outcomes.

School-level Improvement Approach and Implementation Quality

Our second research question asks about the extent to which features of the improvement approach were linked to implementation quality and desirable student outcomes. A summary of our findings related to this question is displayed in Figure 3. The level of the school-level improvement features and implementation quality varied across the three innovation schools, although schools with higher ratings on the three features of the improvement approach also had greater depth of implementation. Implementation quality, however, was not associated with the quantitative outcomes. We first describe the qualitative findings around improvement approach and implementation quality.

Williams High School. Overall, the SOAR team and other school staff at Williams High School had the strongest understanding of the improvement theory underlying the SOAR innovation of the three schools, although it decreased slightly in the second year of implementation. The SOAR team at Williams succeeded in developing an expertise over the SOAR innovation and, unlike the other two innovation schools, developing a shared understanding among school staff of how the innovation practices implemented at their school should lead to student ownership. School staff described SOAR with a high degree of consistency, emphasizing the school focus on goal-setting, grade-monitoring, and growth mindset. In a representative comment summarizing the goals of the innovation at Williams, an English/language arts teacher stated, "Student ownership, to get them to increase accountability for their grades, letting them know where they stand, set goals for themselves to help develop accountable talk when dealing with their grades and school work and everything." This high degree of staff understanding extended to a generally nuanced understanding of growth mindset.

The SOAR team at Williams was also consistently proficient in the use of rapid-cycle testing. Their refinement of the grade monitoring activity through PDSA is characteristic of their proficiency with this element of the improvement process. The focus on grade monitoring was problem-focused and dedicated to improving the form's utility for teachers while also facilitating depth in student responses. In a sequence of multiple, iterative improvement cycles, the team looked at various forms of data, including student and teacher survey data, analysis of student written responses on the form, and outcome data to make evidence-based decisions about how to improve this specific element of the innovation. One SOAR team member reflects, "I feel like we're in the same PDSA cycle, like we've done it kind of back-to-back, but they've been different things back-to-back. So like the first was just implementing grade [monitoring]. That

was our first one in October, and then this one it's still grade [monitoring], but it's a new way we're doing it." This member demonstrates that each cycle leads to the next, thus engaging in continuous improvement. At the end of the year, the team examined patterns in student course failure to connect their grade monitoring activities with student outcomes.

Williams was also consistently proficient in their capacity to engage in the research-practice partnership. Williams had a strong history and culture of cultivating leadership amongst its teachers, and the SOAR team had a track record of successfully leading schoolwide initiatives. One SOAR team member described the team by saying, "a lot of the people on that team are also people on, in other leadership capacities." Through these other teacher leadership capacities, Williams' SOAR team members had experience leading professional development. There is strong evidence that Williams worked productively together to collaboratively achieve their goals, even though there was a lack of an official distribution of tasks and occasional tendency of one member to dominate. In regards to connections with stakeholders outside the team, Williams was intentional about designing messages that would build teacher buy-in and engage early adopters in the work. SOAR members at Williams referenced their prior implementation efforts and regularly drew upon those experiences to assert their position to other teachers as a select group of leaders in the school. Relatedly, the administration saw the teacher leaders as a select group of teachers who were particularly competent to lead the efforts around SOAR. An administrator commented, "this is teacher led, teacher driven…[we are] letting them do what they need to do, because it's working." It was clear that the teacher leaders were seen as a leadership body, with a high degree of leadership capacity, and the deserved autonomy to make decisions and lead SOAR implementation and related professional development.

The proficiency in enacting the improvement approach extended to a moderate quality of implementation in year 1 that was sustained into the second year of implementation. As mentioned above, grade monitoring and growth mindset were the primary practices implemented at Williams. Over three-quarters of teachers in Williams described grade monitoring as the core practice of SOAR. Further, all participants in Williams, including teachers who did not have a second period class and thus did not directly participate, were familiar with the process and its purpose, and all but one teacher reported implementing it every three weeks. While nearly all teachers in Williams implemented the practice, there was variation in the amount of engagement that teachers described having with students during the process. A Williams administrator described this variation by saying that some teachers "just hand out the sheet and say do this... But those teachers who really do engage with their students in conversations about goal setting… then I think they've really gotten a lot out of it." Student comments in Williams are consistent with administrators and teachers in that they indicate nearly all teachers enact the routine, but teachers vary in the depth in which they engage students in it.

A second SOAR component, growth mindset, was enacted strongly in Williams. In the first week of school at Williams, the SOAR team organized the school to deliver a set of seven lessons about growth mindset throughout the day, which all teachers in Williams reported implementing. From the student perspective, they participated in an all-day learning experience around growth mindset. In focus groups, nearly all students in Williams said that they heard about growth mindset "every class the entire day," and students in one focus group called it a "conspiracy." Beyond the first week in school, about the half the teachers in Williams reported also reported using classroom practices that further fostered a growth mindset, such as allowing students to redo assignments or creating an atmosphere where students feel comfortable making

mistakes. Students in Williams confirmed that at least some teachers incorporated ideas about growth mindset, with students in all focus groups providing examples of teachers who continuously reinforce ideas of growth mindset. Despite the inconsistent follow up, most teachers, administrators, and SOAR team members in Williams reported that growth mindset ideas were beginning to be part of the school culture. For example, one teacher said, "even the kids are starting to – it's starting to creep into the vocabulary."

Smith High School. Overall, evidence suggests that Smith High School's enactment of the improvement features were adequate at best. Their team indicated an adequate understanding of the improvement theory, although the level decreased in the second year of implementation and they were never able to achieve a coherent understanding of the SOAR innovation among school staff. When asked about the goals of the SOAR innovation, school staff gave disparate accounts, including keeping up with assignments, problem solving, student discipline, goal-setting, having students overcome obstacles, being productive citizens, college preparation, student empowerment, time management and personal organization, and building student motivation, among others. Even though respondents could not always link the school-wide problem-solving process and behavioral reflection forms to larger goals around student ownership, these practices were generally seen as the main practices they were expected to implement in the first year. In the second year, however, there was more confusion about the practices that comprised the innovation.

We observed a similar pattern in terms of the adequate use of rapid-cycle testing in the first year of implementation with a subsequent decrease in the following year. There was evidence that PDSA was problem-focused, in that the SOAR team drew on student and teacher surveys to improve the school-wide problem-solving process and behavioral reflection forms.

One SOAR team member summarized the utility of their approach: “I think that's where the other teachers are seeing it a little bit better, because we do listen to them and we do want it to work and we want it to be better.” This statement also makes clear that the team was committed to implementing iterative improvement cycles, although, in practice, a specific practice was never tested more than twice. More typical of the team’s engagement in rapid-cycle testing was to focus on a new practice each time, rather than iterate on the same practice.

Most concerning of the school-level improvement features was the absence of demonstrated educator capacity in the first year of implementation, and only limited evidence in the second year. For instance, while the Smith team did accomplish some of their plans, engage in rapid-cycle testing, and lead professional development for their teachers, they needed considerable external support in order to do the work. In contrast to the other schools, Smith struggled to productively collaborate throughout the entire process. One SOAR team member summarized their group by saying they are “not cohesive in what we want to do. We are not organized,” At one point, a team member walked out of a network meeting saying, “I think y’alls goals and my goals are very different and they’re not aligning.” Marked by negativity, conflict and poor attendance at design team meetings, the team was notably demoralized and had little autonomy. However, when a teacher from Williams was assigned as an administrator at Smith, the team began to function better, as described by several members. This member gathered other members for meetings, organized professional learning communities to train teachers, and coordinated the work of the team. While the team and its work might have received an infusion of leadership, other members played supporting roles to the new leader. There was a downside to this approach too. The majority of teachers viewed SOAR as an administrative initiative due to the dual role of leader of the team and administrator that the new member played.

Although Smith was not rated highly for their engagement in the NIC, the quality of implementation was still moderate in the first year of implementation, possibly explaining the positive student outcomes. School stakeholders attributed the depth of implementation that did exist in Smith to the enthusiastic role played by the school administrator who oversaw the improvement efforts. Implementation consisted of a school-wide problem-solving process, use of behavioral reflection forms, and, to a lesser degree, a grade monitoring form. About two-thirds of teachers in Smith described either the problem-solving or behavioral reflection routines as the major emphasis of SOAR. The problem-solving process in Smith, called IPAC (Identify, Plan, Act, Check), was developed after the SOAR team recognized students were having difficulty acting on the goals they had set. One Smith teacher, for example, described the evolution from goal-setting to problem-solving by saying "it's problem-solving and their ability to solve things on their own. … The program was actually implemented where the students put input last fall and then we came up with steps on how to solve problems." Most Smith teachers reported introducing IPAC to students, and about half the students in Smith focus groups indicated the problem-solving steps were a major push in the school and a quarter of the student focus groups described benefiting from IPAC.

Another routine in Smith that most teachers did report implementing frequently: the GROW sheet, which was designed to have students reflect on discipline problems before writing a referral. A Smith administrator described how students frequently fill out GROW sheets, saying "on any given day, almost any period, [I] see at least one student outside of the room working on their GROW sheet. So that means teachers are using them." One Smith teacher explained how the GROW sheet allows the teacher to have individualized conversations with students and allows students to take responsibility for their actions. Students in Smith

appreciated the GROW sheet because it "helps the student and the teacher out…the students won't be able to go to [in-school suspension]…the teacher will talk to the student about what's the problem."

Despite this moderate success with implementation in the first year, the Smith administration decided to withdraw from the SOAR network in the middle of the second year. Some teachers described still engaging in practices like IPAC and the GROW sheet, but without administrative support, implementation declined.

Hancock High School. The Hancock SOAR team demonstrated proficiency in terms of their understanding of the improvement theory, although the depth of understanding did not extend to the majority of teachers in the school. Hancock was distinct in that all SOAR practices were implemented either during weekly advisory lessons or informally through mentoring relationships teachers were encouraged to form with students. Indeed, mentoring and changing the school culture to foster student-teacher relationships became an overarching goal for the Hancock SOAR team. As one Hancock teacher explained it, "This is a great opportunity for a mentorship…Not just the student ownership, but coming in and having a relationship with the child so that they can be some sort of mentor for them." With the goal of changing school culture, there was often less of an emphasis on the specific instructional routines that teachers were expected to change and the practices that were implemented as part of the advisory were often restricted to this weekly lesson. As the advisory periods included practices relevant to student ownership and responsibility and college and career readiness, more broadly, teachers had less of an opportunity to develop a depth of understanding regarding key SOAR practices. Teachers in Hancock frequently spoke about the "SOAR curriculum" and how they were "doing the lessons," with little evidence that SOAR was influencing practice outside the advisory

period, or what the SOAR curriculum was intended to promote. One teacher described how the broader goals of SOAR were not clear, “I don't know that they've ever mentioned the goals to us. …they're working so hard. You would think they're working towards a common purpose.”

The SOAR team’s implementation of rapid-cycle testing was limited. The team took steps to revise the lessons that were implemented during advisory periods, often incorporating feedback from school stakeholders in the process. Yet, these improvements were not done in the context of disciplined rapid-cycle testing. Instead, lessons were revised based on informal teacher feedback rather than being data-driven. The SOAR team did collect teacher surveys about the SOAR lessons and demonstrated concern about how teachers were responding to the lessons, fieldnotes from the design team meetings described little connection between what the team discussed learning from the cycle they just completed to what they now wanted to focus on. As the lessons were developed as part of a year-long curriculum, they were not iteratively improved, as they were only implemented once each year.

The SOAR team at Hancock was consistently proficient in their capacity to engage in the research-practice partnership. Similar to Williams in some ways, the Hancock SOAR team had a good understanding of growth mindset and problem-solving and how they related to the goals of SOAR. The Hancock team was especially strong in their social connections to each other and other teachers in the school. One member explained, “some of us are friends, but more than anything we’ve all been working together for quite a bit, so we have that mutual respect.” The initial human capital at Hancock was not as high as it was at Williams, with fewer team members having experience with activities like leading teacher professional development, but they were able to leverage their social capital along with their human capital in order to lead SOAR in their school. They wanted to build faculty buy-in by getting them involved in the lesson planning and

prototyping. To a large degree, they were able to accomplish this goal. For instance, Hancock demonstrated strong social capital throughout the entire process as they displayed strong working relationships and collaboration and repeatedly emphasized the need to use their relationships with other teachers to obtain stakeholder input on the innovation, thus leveraging their social networks into social capital.

Similar to the other two schools, Hancock had moderate implementation quality in the first year. Yet, the sustained use of the advisory period contributed to a high level of implementation in the second year. Grade monitoring, growth mindsets, and problem solving were all taught in advisories. Yet, as these practices were part of a larger curriculum, the grade-monitoring activity was perceived as less central to the successful implementation of SOAR than at the other two schools. Other advisory topics focused on college and career readiness, such as understanding transcripts, financial aid, and college admissions. Over half of Hancock teachers indicated they have not been asked to do anything outside of advisory. Even though Hancock demonstrated moderate-to-high implementation quality within the advisory period, it is possible that the way these practices were implemented limited the extent to which they had the potential to improve student outcomes in ways observed at the other two innovation schools.

Linking Implementation Quality and Student Outcomes

As evident in Figure 3, the qualitative data on overall quality of implementation does not have a clear association with the quantitative estimates of impact on student outcomes. All schools had moderate implementation in the first year, with Hancock improving implementation over time, Smith declining in implementation, and Williams holding steady. Yet, there is no consistent evidence that SOAR improved student outcomes in Hancock, while there is evidence at both Williams and Smith that SOAR led to some improvements for students. To further

investigate this apparent disagreement between the qualitative and quantitative data, we examined how educators at each school described the impact of SOAR in their school. The qualitative data also provides evidence on how school stakeholders perceived SOAR's impact on student outcomes, which we can triangulate with the quantitative student outcome data. In general, teachers and administrators in all three schools felt the innovation had a palpable influence on students' academic engagement and classroom behavior. In particular, staff at all schools indicated the grade monitoring routine of SOAR has helped students be more aware of their grades. A Williams teacher indicated: "The thing that I really like to see is to see the kids talking to each other about [their grade sheet]. I hear more academic conversations than inappropriate ones … That's something I wouldn't have heard last year." Similarly, a Hancock teacher said, "the biggest focus that I've seen this fall…just getting kids to really be aware of their current status, grade-wise, and how to ask questions about their grades."

Teachers in Williams and Smith also suggested students were not only more aware of their grades, but demonstrated more ownership of their grades by completing assignments and going to tutoring. A teacher in Williams said, "I'm starting to see a little bit, changes in the kids, because they are starting to take more ownership into their learning, and they ask questions that kids in years past wouldn't have asked." This increased awareness of grades stands in contrast to the culture that used to exist in the school, where students did not always link their grades with their class performance. By shifting the locus of control from teachers to students, students were described as taking more ownership over monitoring their performance. Several Smith teachers we spoke to saw a connection between students setting goals to improve their grades and a decrease in the number of incomplete and missing assignments. For example, a math teacher said, "I used to struggle to get my failure rate down to 20%. The last six weeks it was at 8%."

In contrast to the other two schools, fewer Hancock teachers described systematic changes in student academic engagement and classroom behavior. For example, one Hancock teacher described a conversation with a student, "I said, 'Well, have you talked to your teacher?' No…So they don't see that there's a solution to that. They just sit – they would rather fail the class than to go talk to the teacher." Another Hancock teacher said, "At least in conversation, they'll be like, yeah, I can do this to get better. But in practice, it doesn't always go through."

Our explanation for the difference in both perceived and actual impact of SOAR on outcomes across schools is more suggestive than conclusive. However, given the implementation in Williams and Smith focused more on specific routines, such as the grade monitoring, goal-setting, and discipline reflection, while Hancock focused more on implementing lessons in an advisory period, it is possible that the focus on routines led to more change in student behavior outside of the focused SOAR time.

Discussion

Although we found no overall evidence of a relationship between the implementation of the co-constructed SOAR innovation and student outcomes, evidence of outcomes differed across the three innovation schools. Regardless of model specification, students in Williams and Smith high schools, had improved grades and decreased absenteeism, and a reduction in course failure at Smith. At Hancock, we found no consistent evidence of benefits to student outcomes, and even some evidence of a negative effect on student outcomes in the DD model (i.e. increased student absences, decreased average grades, and increased course failure). This variation in student outcomes across school sites is to be expected given the inherent flexibility in an improvement model that privileges adaptive integration, which schools may be more or less able to do successfully (Hannan et al., 2015).

Yet, we also did not find consistent evidence that school engagement in the partnership shaped student outcomes. Generally, schools that were rated higher in terms of the level of the school-level improvement features also had a greater depth of implementation. Yet, neither of these school-level measures was consistently related to desirable student outcomes. Williams was proficient (or better) in terms of their engagement with the NIC, had moderate implementation quality that was sustained over time, and had positive student outcomes. Positive student outcomes were also observed in Smith, although their participation in the NIC was rated as adequate and implementation quality was moderate, both of which decreased over time. Further complicating this story is the lack of consistent evidence of improvements in student outcomes (and some evidence of worsening educational outcomes) in Hancock, but evidence of stronger participation in the NIC than Smith and moderate-to-high implementation quality, depending on the implementation year.

We suggest two main explanations for this finding. First, the qualitative analysis suggests differences in student outcomes may be explained by implementation emphasis in each of the schools. Williams and Smith both implemented practices that were designed to change teachers' instructional routines, with the SOAR team at Williams also establishing a high degree of coherence around the routines. In contrast, at Hancock, the use of advisory periods created more disparate goals for the SOAR innovation, which included an emphasis on improving student responsibility but also improvements to school-wide culture and enhanced student-teacher relationships. Notably, Hancock focused more on cultural and relationship changes, with less emphasis on teachers changing their practice. The data also indicate Williams had the strongest implementation of both routines and sense of coherence around the routines, and the quantitative evidence suggests the most consistently positive outcomes on students. These findings are

consistent with research on other school improvement initiatives that highlight differences between commitment to the idea of the reform rather than to the specific routines the reform requires (Rutledge, Brown, & Petrova, 2017). School improvement requires attending to both the process of change and the specific routines or instructional practices the reform is trying to change; neglecting one over the other may not lead to improved outcomes (Hatch, 2002; Rowan, Correnti, Miller, & Camburn, 2009).

The implications of this for educational improvement are the need to focus on both the process of improvement and the effective practice that represents the improvement. In other words, we need a practice worthy of being scaled and a process that supports successful scaling, even as the mechanisms to achieve those goals may conflict (Rubin, Patrick, & Goldring, 2017). Network-based improvement approaches begin with the recognition that highly structured reforms experience challenges in scaling up because they conflict with existing structures and environments the school faces (Berends, Bodilly, & Kirby, 2002). This flexibility in program design is particularly important for achieving scale; innovations must be able to fit within contexts that vary greatly in organizational structure, buy-in, capacity, and funding while coping with change, promoting ownership, building capacity, and enable effective decision-making (Cohen, Peurach, Glazer, Gates, & Goldin, 2013; Peurach & Glazer, 2012). Network-based improvement models need to find a balance between clearly specifying what fidelity to the core practices involves, while also fostering structured adaptation (Quinn & Kim, 2017). A better understanding of how NICs negotiate this fidelity versus adaptation question is critical, given that the current application of improvement science departs from norms of experimental design.

Our second potential explanation highlights the difficulty of doing this type of network-based improvement. While our mixed methods research design focused on school-level

improvement features linked to observed student outcomes, it could be that there are features of the leadership or organization of the coordinating hub of the NIC that explain the variation in student outcomes across schools. Indeed, the successful launching of a NIC requires careful attention to hub or partnership-level dynamics (Russell et al., 2017). School-level (or teacher-level) variation is the exact problem that educational improvement efforts need to solve (Bryk, Gomez, Grunow, & LeMahieu, 2015). Thus, the differences between schools in their engagement in the NIC and in outcomes may reflect the hub's ability to lead and organize the network. Future research should measure both school-level and hub-level activities in a way that support identifying hub-level activities that facilitate productive school-level processes.

An important implication of this conclusion is the recognition of the need for strong hub-level infrastructure that supports school-level work (Peurach & Glazer, 2012). The nested structure of schools systems requires an infrastructure and shared understandings so that learning reaches beyond a single school (Redding, Cannata, & Miller, 2018). Yet, the examples of these types of successful improvement networks indicates it takes several years for the hub to develop this capacity (Peurach & Neumerski, 2015). Thus our findings may be explained by the fact that the hub itself was in its infancy. Further, developing sufficient infrastructure at the hub-level requires a larger environment in both schools and in educational support organizations that foster improvement-oriented mindsets and processes, yet education has not yet developed these capacities (Peurach, Penuel, & Russell, 2018). Broader improvement at scale depends on developing the support mechanisms that facilitate organizing improvement networks.

There are several potential limitations to this study. One limitation is that some of the outcome variables could be seen as endogenous. Indeed, teachers were responsible for implementing the innovation and assigning grades, determining which students failed, and

writing disciplinary infractions. It is possible teachers may have artificially raised student grades or failed fewer students. Other evidence from the district suggests this is unlikely. In particular, throughout our multiyear partnership we heard about pressures on teachers to not fail students that pre-date the SOAR innovation. Indeed, helping teachers maintain academic press in a context that is focused on increasing graduation rates was one of the findings that led to the focus on student ownership. The endogeneity of outcomes may also be a problem for the number of disciplinary infractions, as the use of the behavioral reflection form before a referral may more directly decrease the number of documented infractions. However, we found little consistent evidence that implementation of SOAR had an impact on the number of infractions.

Another limitation is the possibility that unmeasured school-level processes may also explain differences in student outcomes. While the SOAR teams at Williams and Hancock were comprised of instructional staff, the SOAR team at Smith was led by an administrator, whose leadership over the SOAR team may have made it easier to ensure a high degree of implementation in the first year. Yet, this approach also resulted in a perception that SOAR was an administrator-led initiative, which meant implementation declined in the second year when the administration developed new priorities. Similarly, the different pre-treatment trends in Hancock suggest processes were happening in that school that go beyond the SOAR initiative and that our measures of their engagement in the SOAR improvement process and implementation did not identify.

A third potential limitation is the diffusion of the SOAR innovation beyond the three innovation schools. The SOAR design team did include administrators from two other high schools, each of which indicated some adoption of a few SOAR practices in their schools. While evidence on implementation in these schools is less systematic, the administrators report such

practices were diffused on a small scale, such as to teachers in the department that administrator oversaw. Consequently, our results would then be underestimated if improvements were also in other comparison schools in the district.

Despite these limitations, we believe these findings have practical significance as substantial educational improvement efforts are focused around research-practice partnerships and network-based improvement approaches. First, we provide examples of mixed-methods research conducted in the context of an RPP. School-specific memos about implementation, based on fieldwork visits to the schools, were shared with school and district stakeholders and used to inform subsequent improvement work. These data are important for policymakers and leaders to understand the successes and challenges of implementation. Districts and states often lack resources to collect this type of data, even though they need information on the policy and systemic challenges in implementation, in addition to rigorous estimates of overall impact (Conaway, et al. 2015; Lopez-Turley & Stevens, 2015). Second, this partnership approach is gaining popularity with the assumption that greater attention to the context of implementation will result in more sustained improvements at scale (Means & Penuel, 2005). This study provides evidence that these types of partnerships, while designed around traditional challenges to scaling up, still have difficulties in leading to consistent improvement in student outcomes. The differences in outcomes between schools suggests that this type of improvement process may create opportunities for some schools to do well while others continue to struggle, making the hub or network-level infrastructure critically important. Prior evidence on the importance of having some organizational capacity in order to build more capacity (Hatch, 2002; King & Bouchard, 2011) may mean that continuous improvement approaches to scale may still be challenged in building capacity in all schools.

References

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

Arcia, E. (2006). Achievement and enrollment status of suspended students. *Education and Urban Society*, 38, 359–369.

Berends, M., Bodilly, S., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform: New American Schools after a decade*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/research_briefs/RB8019/index1.html

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

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development*, *78*(1), 246–263.

Bodilly, S. J. (1998). *Lessons from New American Schools' Scale-Up Phase: Prospects for Bringing Designs to Multiple Schools.* (p. 157). Santa Monica, CA: RAND.

Booth, J. L., Cooper, L. A., Donovan, M. S., Huyghe, A., Koedinger, K. R., & Paré-Blagoev, E. J. (2015). Design-Based Research Within the Constraints of Practice: Algebra By Example. *Journal of Education for Students Placed at Risk (JESPAR)*, *20*(1–2), 79–100.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? a review of the predictors of dropping out of high school: precision, sensitivity, and specificity. *High School Journal*, *96*(2), 77.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to Improve: How America's Schools Can Get Better at Getting Better*. Cambridge, MA: Harvard Education

Press.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Cannata, M., & Rutledge, S. A. (2017). Introduction to New Frontiers in Scaling Up Research. *Peabody Journal of Education*, *92*(5), 559–568.

Coburn, C. E. (2003). Rethinking Scale: Moving Beyond Numbers to Deep and Lasting Change. *Educational Researcher*, *32*(6), 3–12. https://doi.org/10.3102/0013189X032006003

Coburn, C. E., & Penuel, W. R. (2016). Research–Practice Partnerships in Education Outcomes, Dynamics, and Open Questions. *Educational Researcher*, *45*(1), 48–54.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). *Research-Practice Partnerships: A Strategy for Leveraging Research for Educational Improvement in School Districts*. New York, NY: William T. Grant Foundation.

Cohen, D. K., Peurach, D. J., Glazer, J. L., Gates, K. E., & Goldin, S. (2013). *Improvement by Design: The Promise of Better Schools*. Chicago ; London: University Of Chicago Press.

Cohen-Vogel, L., Cannata, M., Rutledge, S., & Socol, A. R. (2016). A Model of Continuous Improvement in High Schools: A Process for Research, Innovation Design, Implementation, and Scale. *Teachers College Record*, *116*(13), 1–26.

Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., & Wang, Q. (2015). Implementing Educational Innovations at Scale: Transforming Researchers Into Continuous Improvement Scientists. *Educational Policy*, *29*(1), 257–277.

Conaway, C., Keesler, V., & Schwartz, N. (2015). What Research Do State Education Agencies Really Need? The Promise and Limitations of State Longitudinal Data Systems.

Educational Evaluation and Policy Analysis, *37*(1_suppl), 16S-28S.

Datnow, A., Hubbard, L., & Mehan, H. (2002). *Extending Educational Reform: From One School to Many* (1st ed.). Routledge.

Dede, C. (2006). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. In R. K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 551–566). Cambridge, UK: Cambridge University Press.

Duckworth, A., & Yeager, D. (2015). Measurement Matters Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Durlak, J. A., & Dupre, E. P. (2008). Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *Am J Community Psychol*, *41*, 327–350.

Dweck, C. S. (2007). *Mindset: The New Psychology of Success* (Reprint). Ballantine Books.

Elmore, R. (1996). Getting to Scale with Good Educational Practice. *Harvard Educational Review*, *66*(1), 1–27.

Faria, A.-M., Sorenson, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). *Getting students on track for graduation: Impacts of the Early Warning Intervention and Monitoring Systems after one year* (No. REL 2017-272). Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

Farrington, C., Roderick, M., Allensworth, E., Nagaoka, J., Seneca Keyes, T., Johnson, D., & Beechum, N. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance: A Critical Literature Review*. Chicago: The University of Chicago Consortium on Chicago School Research.

Fishman, B. J., Penuel, W. R., Allen, A.R., & Cheng, B. H. (2013). *Design-based implementation research: theories, methods, and exemplars*. New York, NY: National Society for the Study of Education.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, *74*(1), 59–109.

Fullan, M. (2001). *The new meaning of educational change*. Teachers College Pr.

Hannan, M., Russell, J. L., Takahashi, S., & Park, S. (2015). Using Improvement Science to Better Support Beginning Teachers: The Case of the Building a Teaching Effectiveness Network. *Journal of Teacher Education*, *66*(5), 494–508.

Hatch, T. (2002). When Improvement Programs Collide. *The Phi Delta Kappan*, *83*(8), 626–639.

Henrick, E. C., Cobb, P., Penuel, W. R., Jackson, K., & Clark, T. (2017). *Assessing Research-Practice Partnerships: Five Dimensions of Effectiveness*. New York: William T. Grant Foundation.

Honig, M. I., Venkateswaran, N., McNeil, P., & Twitchell. (2014). Leaders' use of research for fundamental change in school district central offices: Processes and challenges. In K. S. Finnigan & A. J. Daly (Eds.), *Using Research Evidence in Education: From the Schoolhouse Door to Capitol Hill* (2014 edition, pp. 33–52). Cham ; Heidelberg ; New York: Springer.

Jaquith, A. (2017). *How to Create the Conditions for Learning*. Cambridge, MA: Harvard Education Press.

King, M. B. & Bouchard, K., (2011) The capacity to build organizational capacity in schools. *Journal of Educational Administration* 49(6): 653-669.

Langley, G. J. (2009). *The improvement guide: a practical approach to enhancing*

organizational performance. San Francisco: Jossey-Bass.

LeMahieu, P. G., Grunow, A., Baker, L., Nordstrum, L. E., & Gomez, L. M. (2017). Networked Improvement Communities: the discipline of improvement science meets the power of networks. *Quality Assurance in Education*. https://doi.org/10.1108/QAE-12-2016-0084

López Turley, R. N., & Stevens, C. (2015). Lessons from a School District-University Research Partnership: The Houston Education Research Consortium. *Educational Evaluation and Policy Analysis*, *37*(1). https://doi.org/10.3102/0162373715576074

Mac Iver, M. A., & Messel, M. (2013). The ABCs of Keeping On Track to Graduation: Research Findings from Baltimore. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(1), 50–67. doi:10.1080/10824669.2013.745207

Mazzeo, C., Fleischman, S., Heppen, J., & Jahangir, T. (2016). Improving High School Success: Searching for Evidence of Promise. *Teachers College Record*, *116*(13), 1–32.

Means, B., & Penuel, W. R. (2005). Scaling Up Technology-Based Educational Innovations. In C. Dede, J. P. Honan, & L. C. Peters (Eds.), *Scaling Up Success : Lessons Learned from Technology-Based Educational Improvement* (1st ed., pp. 176–197). San Francisco, CA: Jossey-Bass.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook. Second edition*. Thousand Oaks, CA: Sage Publications.

Murphy, J. F., & Torre, D. (2014). *Creating Productive Cultures in Schools: For Students, Teachers, and Parents* (1 edition). Thousand Oaks, California: Corwin.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (Third edition). Thousand Oaks, CA: Sage Publications.

Peurach, D. J., & Glazer, J. L. (2012). Reconsidering replication: New perspectives on large-

scale school improvement. *Journal of Educational Change*, *13*(2), 155–190.

Peurach, D. J., & Neumerski, C. M. (2015). Mixing metaphors: Building infrastructure for large scale school turnaround. *Journal of Educational Change*, *16*(4), 379–420.

Peurach, D. J., Penuel, W. R., & Russell, J. L. (2018). Beyond Ritualized Rationality: Organizational Dynamics of Instructionally-Focused Continuous Improvement. In M. Connolly, D. Eddy-Spicer, C. James, & S. Kruse (Eds.), *The SAGE Handbook of School Organization*. SAGE Publications.

Quinn, D. M., & Kim, J. S. (2017). Scaffolding Fidelity and Adaptation in Educational Program Implementation: Experimental Evidence From a Literacy Intervention. *American Educational Research Journal*, 0002831217717692.

Redding, C., Cannata, M., & Miller, J. (2018). System Learning in an Urban School District: A Case Study of Intra-district Learning. *Journal of Educational Change*, *19*(1), 77–101.

Redding, C., Cannata, M., & Taylor Haynes, K. (2017). With Scale in Mind: NCSU's Integrated Model of School-Based Design and Implementation. *Peabody Journal of Education*, *92*(5), 589–608.

Rowan, B., Correnti, R. J., Miller, R. J., & Camburn, E. M. (2009). School improvement by design: Lessons from a study of comprehensive school reform programs. In *AERA handbook on education policy research* (pp. 637–651). New York, NY: Routledge.

Rubin, M., Nguyen, T., & Cannata, M. (2015). The Influence and Development of Capital for Teacher Leadership. Presented at the University Council for Educational Administration annual convention, San Diego, CA.

Rubin, M., Patrick, S., & Goldring, E. (2017). Make it quick or make it last? Dilemmas of prescriptive practices and perceived alignment in program implementation. *Peabody*

Journal of Education, *92*(5), 609–626.

Russell, J. L., Bryk, A. S., Dolle, J. R., Gomez, L. M., LeMahieu, P. G., & Grunow, A. (2017). A Framework for the Initiation of Networked Improvement Communities. *Teachers College Record*, *119*(7).

Rutledge, S. A., Brown, S., & Petrova, K. (2017). Scaling Personalization: Exploring the Implementation of an Academic and Social-Emotional Innovation in High Schools. *Peabody Journal of Education*, *92*(5), 627–648.

Schneider, M. (2018, July 30). Developing an Evidence Base for Researcher-Practitioner Partnerships. Retrieved October 5, 2018, from https://ies.ed.gov/blogs/director/post/rpps

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*(3–4), 207–231.

Siskin, L. S. (2016). Mutual Adaptation in Action. *Teachers College Record*, *118*(13), 1–18.

Smith, T.M., Cannata, M., & Taylor Haynes, K. (2016). Reconciling data from different sources: Practical realities of using mixed methods to identify effective high school practices. *Teachers College Record 118*(7): 1-34.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. *Review of Educational Research*, *72*(3), 387–431. https://doi.org/10.3102/00346543072003387

Sowers, N., & Yamada, H. (2015). *Community College Pathways: 2013-2014 Descriptive Report*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Supovitz, J. A. (2008). Implementation as iterative refraction. In J. A. Supovitz & E. H. Weinbaum (Eds.), *Implementation Gap: Understanding Reform in High Schools* (pp. 151–172). New York, NY: Teachers College Press.

Thompson, M., & Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (pp. 1–44). Princeton, NJ: ETS.

Tichnor-Wagner, A., Wachen, J., Cannata, M., & Cohen-Vogel, L. (2017). Continuous improvement in the public school context: Understanding how educators respond to Plan-Do-Study-Act cycles. *Journal of Educational Change*, *18*(4), 465–494.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, *13*(1), 12–28.

Tseng, V. C., & Nutley, S. (2014). Building the Infrastructure to Improve the Use and Usefulness of Research in Education. In K. S. Finnigan & A. J. Daly (Eds.), *Using Research Evidence in Education: From the Schoolhouse Door to Capitol Hill* (pp. 163–176). New York: Springer.

Tables and Figures

Figure 1. Conceptual Framework of How Networked Improvement Communities Shape Implementation and Student Outcomes

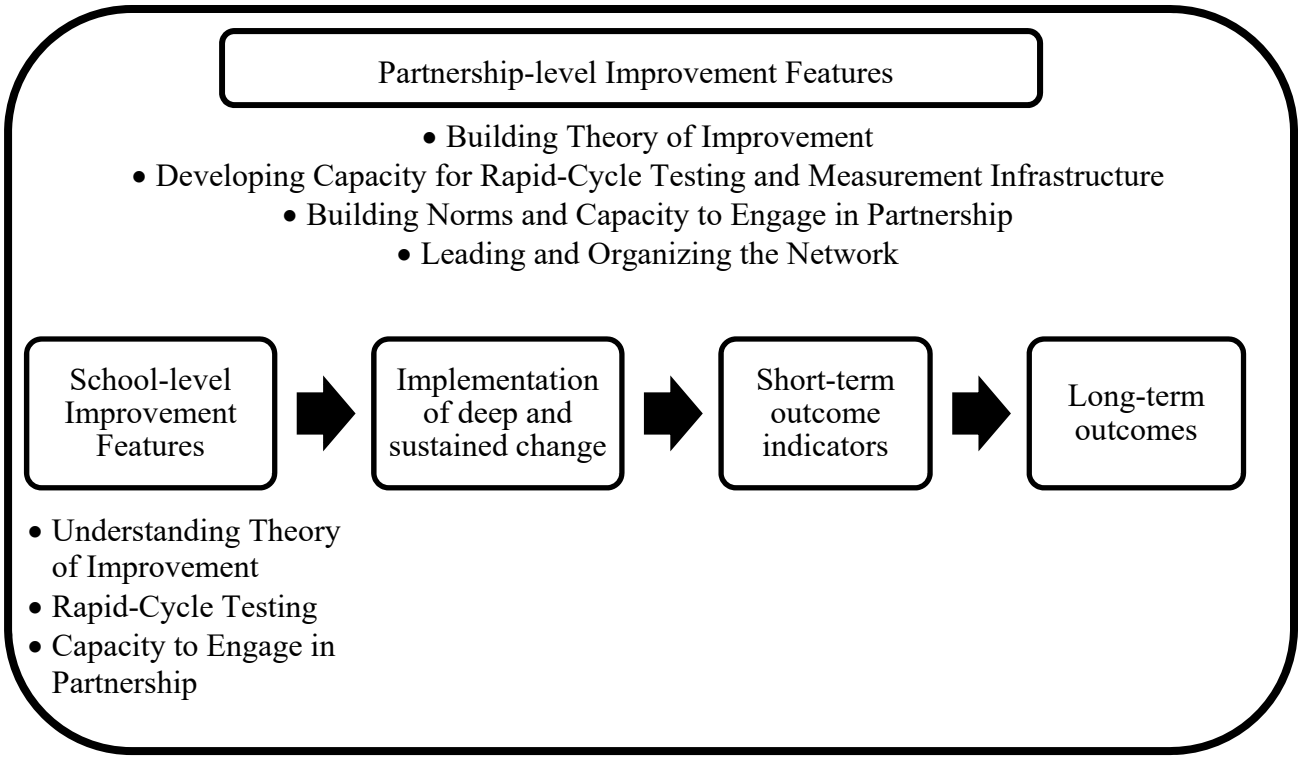


Figure 2. Pre- and Post-Treatment trends of Student Passing Rates, Average Grades, Attendance, and Number of Infractions

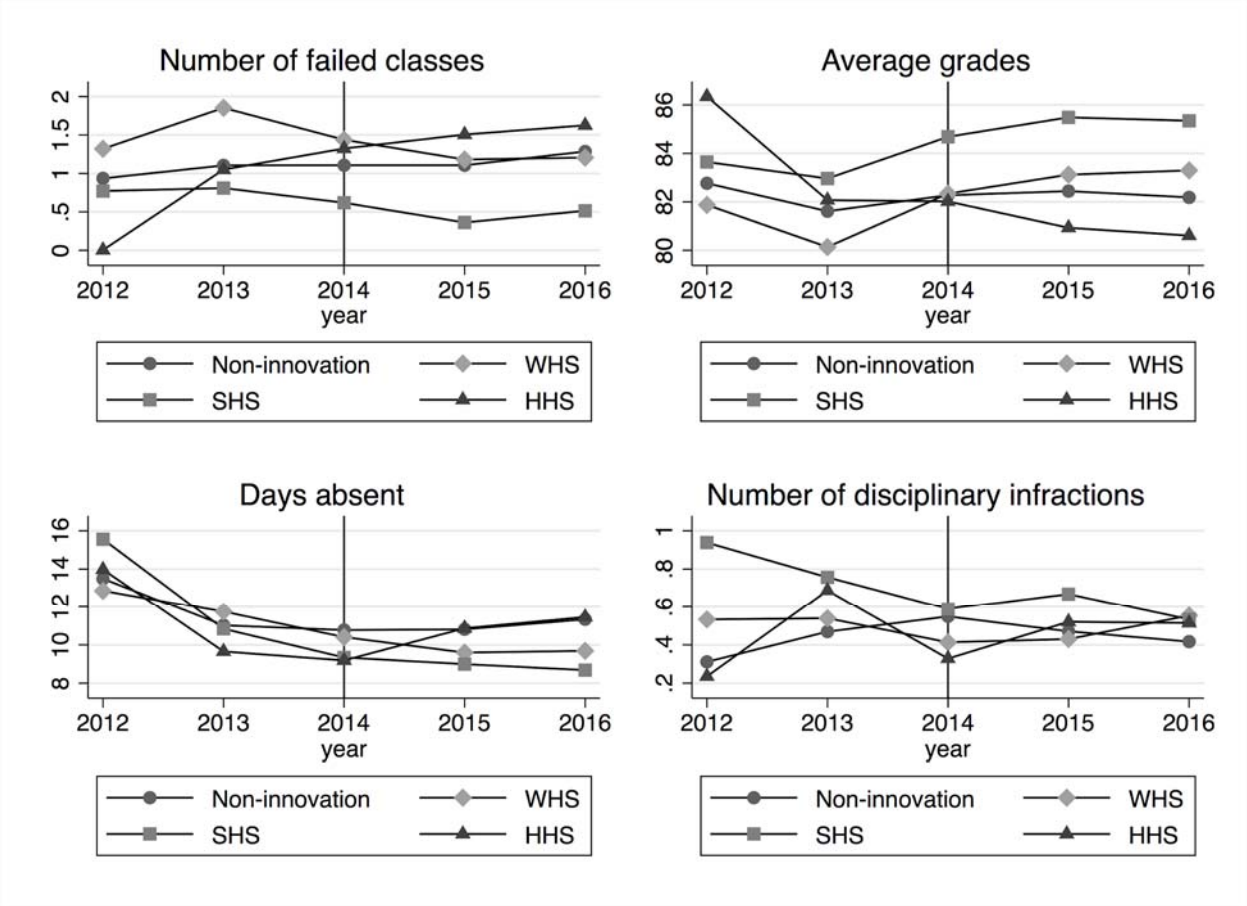
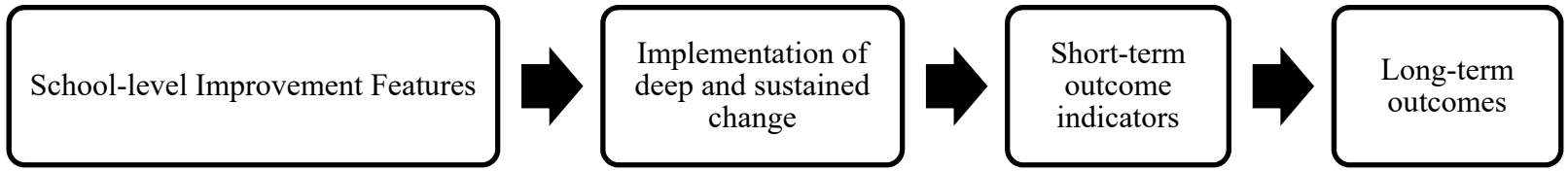


Figure 3. Evidence of How Networked Improvement Communities Shape Implementation and Student Outcomes



Williams H.S.	•Excellent understanding of improvement theory that decreased slightly over time •Proficient use of rapid-cycle testing •Proficient partnership capacity	Moderate implementation quality sustained over time	Decreased student absenteeism	Improved student grades
Smith H.S.	•Adequate understanding of improvement theory that decreased slightly over time •Adequate use of rapid-cycle testing, but decreased use over time •Initial absence of partnership capacity, but improvements over time	Moderate implementation quality decreased over time	Decreased student absenteeism	Improved student grades and decreased course failure
Hancock H.S.	•Proficient understanding of improvement theory •Limited use of rapid-cycle testing •Proficient partnership capacity	Moderate initial implementation quality improved over time	No consistent evidence of short-term improvements	No consistent evidence of long-term improvements

Notes. School-level improvement features rated on a five-point scale: Excellent, Proficient, Adequate, Limited, and Absent. Implementation of deep and sustained change rated on a three-point scale: High, Moderate, and Low.

Table 1 – Timeline of Partnership Activities

Phase	Time period	District design team activity	School SOAR team activity	Research activity
Design and Development	Winter/ Spring 2013	Developed initial innovation prototype; Monthly two-day meetings	n/a	Observation of design team meetings
Piloting	2013-14	Monthly meeting to oversee PDSA cycles and work with SOAR teams to develop innovation	Monthly two-day network meetings; Engage in PDSA cycles to develop the innovation; Initial teacher professional development; Biweekly check-in meetings	Lead PDSA trainings and facilitate cycles; Observations of network meetings; One research visit to each school
Full implementation	2014-15	Quarterly meetings to plan for scale out and sustainability	Initial implementation of fully developed innovation; Continued engagement in PDSA cycles; Quarterly network meetings to share learning; Monthly check-ins; Teacher professional approximately monthly	Support PDSA cycles; Two research visits to each school; Observations of network meetings
Scale out	2015-16	District offices gradually assumes responsibility for facilitating network and supporting work in schools; Quarterly meetings; Four schools join network	Year 2 of full implementation in innovation schools; Continue to engage in PDSA and share learning in quarterly meetings; Professional development as necessary	Support PDSA cycles; One research visit to each school; Observations of network meetings

Table 2. Descriptive Characteristics Prior to Implementation

	Non-Innovation Schools	Innovation Schools (all)	Williams	Smith	Hancock
Number of failed classes	1.10	1.07	1.25	0.74**	1.35*
Average grade	82.22	83.03***	83.19*	83.84**	81.28*
Days absent	11.36	9.98***	10.39	9.91*	9.32**
Number of disciplinary infractions	0.58	0.45***	0.40	0.58	0.33
Free or Reduced Price Lunch	0.69	0.64***	0.44***	0.74	0.84*
Black student	0.25	0.14***	0.20	0.12*	0.05**
Hispanic student	0.59	0.69***	0.46	0.80*	0.92***
Other race	0.04	0.03***	0.04	0.03	0.01**
Gifted	0.12	0.12	0.17	0.07*	0.11
Days enrolled	169.42	169.70*	170.74	168.73	169.45
Withdrew	0.13	0.13	0.16	0.11	0.13
Late start	0.10	0.10	0.10	0.09	0.09
Number of Courses	13.21	12.85***	12.99	12.89	12.52**
Fraction of Black students	0.25	0.14	0.21	0.12	0.05*
Fraction of Hispanic students	0.58	0.68	0.45	0.79*	0.92**
Fraction FRPL	0.62	0.59	0.40**	0.68	0.79**
School size	1766.30	1740.51	2010.00	1859.00	1016.00
Observations	14406	4439	1798	1695	946

Notes. *t*-test of significant differences accounts for school-level clustering. Descriptive statistics reported for 2013-2014 school year. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table 3. Estimates of the Effect of the Innovation on Student Passing Rates, Average Grades, Attendance, and Number of Infractions

	Days absent	Number of disciplinary infractions	Number of failed classes	Average grades	Days absent	Number of disciplinary infractions	Number of failed classes	Average grades
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Innovation school	-1.17*	0.08**	-0.21	0.95+	-0.77	-0.02	-0.23	0.67
	(0.46)	(0.02)	(0.15)	(0.50)	(0.64)	(0.08)	(0.20)	(0.82)
Constant	-5.04***	-0.10	0.22	17.45***	5.36***	0.30	0.32+	77.60***
	(0.98)	(0.18)	(0.16)	(1.41)	(0.88)	(0.20)	(0.16)	(0.99)
Lagged dependent variable	x	x	x	x				
Year Fixed Effect	x	x	x	x	x	x	x	x
School Fixed Effect					x	x	x	x
Observations	60456	62408	58817	58811	85680	85680	85680	85673
R-squared	0.41	0.27	0.27	0.54	0.12	0.12	0.11	0.21

Notes. All models control for FRPL, student race/ethnicity (Black, Hispanic, other race), gifted status, days enrolled, number of courses, grade level, and indicators if the student started school after the beginning of the school year or withdrew before the end of the year. Robust standard errors clustered at the school level in parentheses. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table 4. Estimates of the Effect of the Innovation on Student Passing Rates, Average Grades, Attendance, and Number of Infractions, by Innovation School

	Days absent	Number of disciplinary infractions	Number of failed classes	Average grades	Days absent	Number of disciplinary infractions	Number of failed classes	Average grades
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Williams HS	-1.25**	0.05*	-0.11	0.74*	-1.05*	0.04	-0.45***	1.42**
	(0.38)	(0.02)	(0.08)	(0.30)	(0.46)	(0.05)	(0.10)	(0.47)
Smith HS	-1.10*	0.10***	-0.44***	1.64***	-1.36*	-0.12*	-0.33**	1.24*
	(0.50)	(0.02)	(0.08)	(0.35)	(0.45)	(0.05)	(0.10)	(0.46)
Hancock HS	-1.17+	0.11***	0.12	-0.32	1.15*	0.08	0.49***	-2.26***
	(0.54)	(0.02)	(0.10)	(0.38)	(0.47)	(0.06)	(0.09)	(0.43)
Year Fixed Effect	x	x	x	x	x	x	x	x
Lagged dependent variable	x	x	x	x				
School Fixed Effect					x	x	x	x
Observations	60456	62408	58817	58811	85680	85680	85680	85673
R^2	0.41	0.27	0.27	0.54	0.12	0.12	0.11	0.21

Notes. All models control for FRPL, student race/ethnicity (Black, Hispanic, other race), gifted status, days enrolled, number of courses, grade level, and indicators if the student started school after the beginning of the school year or withdrew before the end of the year. Models include grade fixed effects. Robust standard errors clustered at the school level in parentheses. + p<0.10; * p<0.05; ** p<0.01; *** p<0.001.