Continuous Improvement in Action: Educators' Evidence Use for School Improvement

Christopher Redding
Marisa Cannata
Mollie Rubin
Vanderbilt University

Abstract
Continuous improvement approaches to school reform involve the use of data to understand a problem and test and refine potential solutions. In this paper, we describe how educators come to understand and use data for continuous improvement within a large, urban district. We find evidence that educators are likely to draw on a mix of evidence as well as evidence substitutes when refining the innovation at the school-level. While implementation teams considered outcome data, they gravitated towards perceptual data. They used this data to make small improvements to the innovation but also to gauge the level of teacher buy-in and make modifications that would better meet teachers’ needs.

Acknowledgment: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E100030. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.
Introduction

With underwhelming results from a research-to-practice paradigm that seeks to replicate evidence-based practices in new sites (Bryk, Gomez, Grunow, & LeMahieu, 2015), continuous improvement research is rising in prominence. Continuous improvement research is based on the idea that taking effective practices to scale requires an understanding of the conditions and contexts that enable program success (Cohen-Vogel et al., 2015). With a greater sensitivity to how local contextual factors shape the impact of educational interventions, a key piece of continuous improvement is the use of multiple forms of evidence. Evidence is drawn on not only to understand whether or not an intervention is achieving positive results, but also to understand how local conditions are producing this outcome (Lewis, 2015).

Much of the research on educator evidence use has gravitated towards the obstacles to successful use, including the lack of capacity to select relevant evidence and interpret results (Mandinach, 2012; Marsh, Pane, & Hamilton, 2006), lack of time (Ingram, Louis, & Schroeder, 2004), and insufficient access to data (Coburn, Honig, & Stein, 2009; Lachat & Smith, 2005; Marsh et al., 2006). This research has been less attentive to the ways in which educators apply their learning and how they assess the ongoing effectiveness of these decisions (Marsh, 2012). Yet using data for assessing effectiveness of interventions and revising them is critical to improvement science. Further, most research on data use focuses on teacher data use for instructional improvement in their own classrooms (e.g. Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006; Simmons, 2012), while improvement science asks educators to use data to make decisions on interventions that impact overall systems and not just classroom instruction.

This paper extends our knowledge of educator data use by reporting on the use of evidence for whole-school improvement by three teacher-led teams. In collaboration with
researchers and program developers, teacher-led teams were engaged in a continuous improvement reform model designed to develop and refine practices to increase student ownership and responsibility (SOAR) within their schools. To explain how educators came to understand and use data for continuous improvement, we ask the following research questions:

1. What were the barriers educators faced as they try to use evidence to improve a school-wide reform?
2. How did educators use evidence for the improvement of a school-wide reform?
3. What types of evidence were they citing to make claims about what they were learning?

In the next section, we review the role of evidence in continuous improvement research. We then situate this paper in the literature on educator data use, with particular attention to the types of evidence educators draw upon and the barriers to using data. Next, we describe the district context of the study, data collection, and method for data analysis. Then we review our results, focusing on the context of evidence use, the barriers to evidence use, and how educator teams used data for improvement. We conclude by reviewing our central findings and linking our results back to continuous improvement research and educator evidence use.

**Improvement Science and Educator Evidence Use**

Recent scholarship on school reform has highlighted the need to use new approaches to enacting change and to recognize the critical interplay between the reform practices, the process of implementation, and the local context (Penuel, Fishman, Cheng, & Sabelli, 2011). One such approach, improvement science, focuses on using continuous improvement processes to address problems of practice in ways that contribute to organizational learning (Bryk et al., 2015). Key to improvement science is the creation of networked improvement communities that engage in
disciplined cycles of inquiry where data is used to understand the problem and test potential solutions, leading to greater understanding through data use (Bryk, Gomez, Grunow, & Hallinan, 2011). Two guiding principles inform the work of improvement communities. These communities seek to address a shared problem of practice. Addressing this problem requires designing a solution that aligns with the problem as it is experienced by local stakeholders (Bryk et al., 2015). Bryk and colleagues (2015) have summed up these two points as being problem-specific and user-centered.

To develop systematic solutions, the improvement process described in this paper uses Plan, Do, Study, Act (PDSA) as a tool for continuous improvement. PDSA is drawn from improvement science as a tool for organizational improvement (Langley et al., 2009). PDSA provides a systematic tool to answer the question: How do we know a change led to an improvement? Answering this question relies on linking a change idea in the design with easily measured outcomes. In addition to outcome data, design team members also collect feedback data about the usability of the practices. These different forms of evidence are used to improve the design and align it with existing school processes before testing it again. In summary, PDSA is used to make evidence-based decisions of how to adapt the design to better fit unique school context.

Using data for school improvement is not new. Indeed, data-driven decision-making has been a priority for 20 years (DeBray-Pelot & McGuinn, 2009). Yet, teacher data use is typically characterized as guiding teachers to make well-informed instructional decisions (Hamilton et al., 2009). Even when teachers are able to use data in ways that benefit their instruction, the data is usually comprised of data directly from the district’s accountability systems. In other words, the
data teachers are asked to draw on for improvement purposes may be used by administrators or central office staff to hold them accountable.

Improvement science notes a critical distinction between data for accountability and data for improvement, with different data needs and social conditions of use (Yeager, Bryk, Muhich, Hausman, & Morales, n.d.). While data for accountability focuses on outcomes to identify whether expectations are being met, data for improvement is more practical. It provides evidence about the practice itself and is grounded in daily practice (Yeager et al., n.d.). This broader focus on evidence better captures the ways in which educator decision-making is not solely based on data, but a range of evidence.

At present, there is a dearth of research on the use of evidence in the context of continuous improvement research. In a special issue of the *Journal of Teacher Education* on the role of improvement science and teacher professional development, researchers review the role that measurement has played in their work (LeMahieu, Edwards, & Gomez, 2015). In one study on the Building a Teaching Effectiveness Network (BTEN), Hannan and colleagues (2015) allude to the challenges faced by the majority of schools in the network when using data for improvement efforts. Eight of the schools in the network struggled with documenting and acting upon what they learned. Two schools were more capable of applying improvement methods. They write, “What set these two schools apart was their use of data and disciplined inquiry to amass knowledge of their school systems and their teachers’ needs, coupled with their strategic application of this knowledge in their support efforts” (p. 502). Although educators may be able to use evidence from local improvement barriers, they face a number of barriers in the process.

**Existing Research on Educator Use of Data**
While improvement science asks educators to use data in different ways, prior reform efforts have sought to increase educator data use and a substantial amount of research exists on the process educators used to interpret data to turn it into usable knowledge (Honig & Coburn, 2008). This research has focused on the types of evidence educators draw upon and the barriers educators face when using data.

*Types of Evidence*

We examine the types of evidence educators draw upon, its perceived relevance, and the social context in which the evidence was examined. Evidence includes data related to student or school outcomes (passing rates, disciplinary incidents), inputs (student demographic information), school processes (data on quality of instruction or program implementation), and perceptions (surveys or focus group discussions) (Marsh, 2012). Educators also draw on evidence substitutes, — anecdotal information, experience, intuition, ideological preference, and customary practice (Bryk & Gomez, 2008; Dormann & Binnewies, 2015; Ingram, Louis, & Schroeder, 2004). In their review of evidence-based decision-making among district central office staff, Honig and Coburn (2008) find district leaders to rely on multiple forms of evidence to ground their decisions. This evidence includes student and school performance data but also feedback from teachers and input from students, parents, and community members. In reference to this local knowledge, they write, “These forms of evidence seem important to central office support for school improvement efforts (p. 586). Considering this diversity of evidence enables us to examine how teacher-led implementation teams come to learn about evidence that could be used for their local improvement efforts.
Barriers to data use

The research on teacher data use has highlighted a number of barriers to effective use of data. These barriers include the lack of access to data (Coburn et al., 2009; Lachat & Smith, 2005; Marsh et al., 2006), the perceived validity of data (Ingram et al., 2004; Marsh et al., 2006), time to engage in data-based inquiry (Ingram et al., 2004), and the capacity to analyze and develop actionable solutions from the data (Mandinach, 2012; Marsh et al., 2006; Supovitz & Klein, 2003). The organizational culture and leadership within a school also shapes the use of data to inform school and instructional improvement (Marsh et al., 2006; Schildkamp & Poortman). Administrators can create school-wide norms for data use (Halverson, Grigg, Pritchett, & Thomas, 2007), while ongoing collaboration fosters the use of data to improve instructional practice (Marsh, 2012).

The broader social context of data also shapes its usage, and possibly creates barriers to using data for instructional and school improvement. The literature is mixed over the extent to which accountability-driven district culture is conducive to or inhibitive of educators’ use of data. Marsh and colleagues (2006) suggest that accountability pressures have increased the availability of data and educators’ capacity to work with data. Accountability pressures may also motivate the active use of data, possibly in efforts to identify and support “bubble kids” (Booher-Jennings, 2005; Jennings, 2012). Besides adding pressure for educators to use data to inform improvements, the culture of accountability may undercut the use of data for instructional and school improvement when evidence of school outcomes is the same data that is used for teacher and school accountability.

Data and Methods
The data used in this paper were collected as part of a larger study in a large urban district in the southwestern United States to explore new approaches of scaling effective practices in high schools. At the time of the study, the district served approximately 80,000 students, the majority of whom are low-income or from traditionally underserved racial/ethnic groups. Approximately a quarter of all students in the district are classified as Limited English Proficient (LEP). The recent history of high school reform prior to this work includes the alignment of locally developed curriculum with academic standards and common assessments across schools and the creation of thematic foci for each high school and student choice in selecting schools based on chosen program. Recent changes to the state high school assessment program also influenced what was occurring in high schools. For the subjects tested on state assessment, the district has introduced benchmark content-based assessments (CBAs) every six weeks. These assessments form the primary basis of conversations around data within schools in the district, as the district has trained administrators and teachers in strategies for reviewing these assessment results and other forms of student data, including student passing rates, attendance, and discipline.

Demographic characteristics of the study schools are reported in Table 1. Pseudonymous names are used to uphold the confidentiality of the participants. Demographic characteristics are rounded to prevent the use of identifying school characteristics. Desert Grove High School (DGHS) enrolls approximately 2,000 students, over half of whom are Hispanic of African American. Forest Glen High School (FGHS) has approximately 1,000 students, 90 percent of whom are Hispanic. Valley High School (VHS) similarly enrolls 1,500 students. 80 percent are Hispanic and another 10% are African American. Selecting schools for this study was a collaborative process with district personnel and school administrators, with the selected schools
expressing an interest and willingness to participate in this innovative reform model. Schools were generally identified as moderately performing schools in the district.

Case Study Data Collection

The data used for this analysis come from two sources. Table 2 contains detailed information about the frequency with which these data sources were collected. We refer to the primary data source as the “process data” and it includes observational and artifact data from district-wide design team meetings, including observational fieldnotes and audio recordings of all design team meetings, artifacts distributed or produced at the meetings, and feedback forms completed by members after each meeting. Audio data from each session were not transcribed, due to the length and relative complexity of each session recording, in which recorders captured discussions among an often dispersed groups of participants. Instead, graduate students listened to each recording in its entirety, and utilized reflection forms to partially transcribe and synthesize data according to the analytic framework described above. Research team members collected these data sources by attending regular meetings with the school implementation teams (i.e., teacher-led teams) and program developers from Spring 2013 to Spring 2015. During Design and Development work (Phase 1 and Phase 2), 12 sessions were held throughout the school year. During Implementation (Phase 3), six sessions were held throughout the year.

The purpose of these sessions shifted throughout the project. In Phase 1, the design team learned about the research that informed the design and collected and analyzed data from the innovation schools to begin the process of translating broad design principles into a design prototype that could be implemented in three schools across the district. To engage in this ambitious work, the team engaged in a series of capacity-building activities to support their leadership of development and implementation efforts in their schools. During the development
phase (Phase 2), these sessions still involved training activities but also served as a collaborative workspace where teams could share their experiences with adapting the design to their school context. Teams began using data to identify the needs in order to customize the emergent design for their students and teachers. More active data use occurred during the implementation phase (Phase 3) as teams drew on a wider array of data to document initial successes related to implementation, both in terms of teacher buy-in, student responsiveness, and student outcomes. PDSA was introduced in Phase 2, with explicit training and coaching in using PDSA during this time. PDSA continued to be applied in Phase 3 during the school-wide implementation.

To supplement these observations from the district meetings, the research team conducted in-depth, semi-structured interviews with participants. Design team members and a subset of members from the school development team were interviewed in the Summer 2013, Summer 2014, and Fall 2015 to understand their experiences with the work. We also conducted fieldwork visits in each of the partner schools to monitor implementation in Phases 2 and 3. Three visits took place between December 2013 and April 2015. For the first visit, teams of nine researchers visited each of the innovation schools for one day. For the second two research visits, teams of three researchers spent four days in each of the schools on each of these visits. Except in the case of turnover, we re-interviewed implementation team members during both visits. In total, we conducted 66 semi-structured interviews with implementation team members. From these interviews, we focus on their descriptions of their experiences with data use and the continuous improvement process.

Data Analysis

A qualitative case study design (Lincoln & Guba, 1985) allows us to probe how school stakeholders, particularly the teacher-led teams, used evidence in the ongoing development of the
innovation. As part of the larger project, process data were analyzed in multiple stages. The process data was analyzed for the design, development, and implementation phases, respectively, to produce summary memos that describe evidence from that phase according to a framework drawn from the project’s framework for innovation design and development. This framework consisted of several *a priori* codes in addition to codes that emerged inductively from the data. From the process analysis, we draw on codes for piloting and Plan, Do, Study Act (PDSA)—the particular model organizing the continuous improvement cycles), participant’s understanding of the continuous improvement process, and the integration of the research team.

Regarding the fieldwork data, the research team coded transcripts based on a coding schema related to the project’s framework for quality implementation. The coding process was iterative in nature with members of the research team comparing coding to ensure a consistent understanding and application of codes (Corbin & Strauss, 2008). The team then wrote a summary memo for each field visit, probing similarities and differences across the three partner schools. For this paper, we draw on analytic codes related to the implementation teams’ use of evidence, specifically in the context of the continuous improvement of the innovation design. From the analysis of the school-based fieldwork, we draw on codes for continuous improvement, the capacity of implementation team members, and external support to the implementation team.

**Results**

*Supporting PDSA*

Before discussing the ways in which educator teams used evidence for their school improvement efforts and the barriers faced as part of this process, we provide an overview of the introduction of PDSA, as it was the main vehicle for evidence use on the school development teams. To engage in the ambitious PDSA process, external facilitators adopted practices
conducive to successful evidence use. These include technical training (Marsh, 2012), cross-school collaborative opportunities (Chen, Heritage, & Lee, 2005; Huffman & Kalnin, 2003; Murnane, Sharkey, & Boudett, 2005), the creation of data teams (Schilkamp, Portman, & Handelzalts, 2015), and the establishment of norms for data use (Ikemoto & Marsh, 2007; Park & Datnow, 2009). At the beginning of Phase 2, PDSA was introduced as a tool PDSA to further develop to innovation prototype and align it to the school’s unique context. When PDSA was first introduced, participants were given a general overview of this approach to continuous improvement that emphasized the role in testing practices and gathering feedback as part of the ongoing and iterative improvement of the innovation design. The enactment of PDSA was highly scaffolded with the facilitators and research partners coordinating the first rounds of PDSA testing. Participating teachers were asked to pilot the same lessons on growth mindset and problem solving. The facilitators provided the timeline for when it should be implemented, told the teams how many teachers should pilot, designed the data collection process, and led the debriefing sessions with school design teams. During this initial phase of PDSA, measures for students and teachers were co-developed with the research team, administered by school design team members, and compiled and analyzed by the research team.

In the spring of the Phase 2, facilitators and researchers continued to scaffold the PDSA process, although they now supported schools in planning and testing their own change ideas. They worked with the school design teams as they planned their work and helped them link outcome measures with their change idea. As school teams came to take on more responsibility over managing their own continuous improvement process, the aims of PDSA broadened as a tool to make systematic improvements to teacher and school practices as well as build staff buy-in. When school teams designed data collection instruments and collected and analyzed their
own data, two key distinctions were made from previous evidence use. First, school development teams were encouraged to collect easy-to-gather data that was closely linked with their change idea. Second, they were urged to school process data on program implementation or perceptual that could help teams understand how the innovation design should be modified. They were also asked to gather outcome evidence of changes in student behavior.

Barriers to Using Evidence for School Improvement

Research on educator evidence use has identified a number of barriers to using data for instructional and school improvement. We focus on two barriers: (1) educator capacity to convert broad innovation design into specific outcome measures and (2) the perceived validity of the data. Beginning in Phase 1, discussions of how to measure outcomes of the SOAR innovation were challenging for educators. Teams struggled to link specific outcomes with the general innovation prototype. As the researchers and program developers shared initial drafts of measures and asked for design team members help in clarifying them, a couple of challenges emerged. First, the design team members struggled to disentangle the practices of the innovation and the outcomes they wanted to reach. When outcomes were discussed, team members had difficulty distinguishing between the change ideas of the innovation, implementation indicators, and proximal and distal outcomes. Without a fully specified innovation design, some participants thought they should focus on the outcome and engage in backward mapping to identify the change that would be implemented. For instance, Frank comments,

> What do we want to see and how are we going to measure it. Because to me, when I’m talking, that really determines what we’re going to ask you to do, right? Because if want to have something measureable, you know, how do we measure that? Is it the amount of homework that got turned in? That determines what we want to be asking teachers to do.

For others on the team, it was more important to first clarify the design itself and then specify the student outcomes they would expect to change as a result. This segment of the design team
describes framing the “scope and sequence” and then linking it with outcomes. One of the more vocal members on the team, Matthew, advocated for this approach: “I don’t think we can make a prototype that is going to guarantee. We can’t design something, I’m feeling like we’re designing something trying to guarantee a result.” He elevated teacher practices and attitudes over student outcomes, as they could more easily monitor whether the SOAR innovation was changing teacher behaviors. He continues,

> We want our schools to be better. Moving forward. All boats rise with the tide. If we get some things going and if we’re trying to be so specific about this is the specific thing we’re trying to improve. I mean, I’m going to know if our implementation went well if I can go and ask a kid, “What did you do this year?” and the kid says, “Well this year was different because we did blah, blah, blah….” Even if it didn’t produce the other outcomes, I’ve changed the behavior.

At the end of Phase 1, the difficulty educators had in converting the broad innovation design into specific outcome measures held up the process of defining how to measure SOAR. A larger, and unrelenting, barrier to using evidence related to deeper concerns of validity of any measures that would be used to measure SOAR.

There was a recurrent hesitancy among team members about converting elements of the innovation into measureable outcomes. For some, this hesitancy was related an underlying skepticism over the validity of any measurement of SOAR. These concerns grew out of concerns with the quantification of teacher practice, particularly in the context of a district with a history of using a variety of metrics for teacher and school accountability. During an interview, Matthew expressed his concern with measurement: “The growth mindset part of it for me is something that I struggle with and have for some time, just because how do you actually change someone’s mindset and then see the results of that in—in a quantitative measure?” For others on the design team, the problem was the inability to attribute any outcomes with SOAR or other school or
district programs. Brett recognizes the multitude of factors that may explain the improvements in passing rates that the team has observed. She says,

Yeah. [SOAR is] not the only factor, especially because last year was the year where they really pushed on pass-fail rates for teachers. They were really being called on the carpet if they had a high fail rate. So this year it is entirely possible that passing rates have increased not just because of the innovation, but because teachers just don't want to deal with it.

While some members expressed fundamental validity concerns of how SOAR would be measured, the dominant concern related to use of data in the district in the past.

The concern over the use of student outcomes related to the ways in which data from campuses had been distorted in the past. These concerns were most apparent during the transition between Phases 1 and 2. Design team members felt that the district would misrepresent their work, decontextualize the innovation from its school-based context, and distort the underlying goals of the innovation in the process. Mark, a central office staff member, summarizes this sentiment:

When we take something that should be going on in class that is very instructional and about relationship building, and conversations, when we boil it down to this, leaders will just look at the data report and not go into the class and care about the conversations and the instructions that is happening. We’re boiling down a human-centered objective into a data point that will be hung around necks as an albatross at some point to collect data, to compile data and to say it’s not being effective.

There was a related suspicion of the data collection plans suggested by other members on the design team. Describing ways to observe the classroom implementation of SOAR, Paulette describes how the district has previously supported departmental learning walks as a way to document the types of materials that other teachers had put on the walls in their classrooms. A conversation between the two teacher leaders at Forest Glen reflects concern over the ways in which administrators and other members in the district use this type of low-stakes data to misrepresent teacher practice:
Clara: I think that part of the problem that we’re having is that we haven’t figured out ways to monitor, to assess, that what is working is working, and I think part of the issue with lots of the district initiatives we’ve had in the past. Let’s use learning walks. How do we know that learning walks are working? I’ll say that on our campus this year we had a serious conversation about learning walks.

Rebecca: You can you have wicker posters, and your pictograms, and your word wall but, cool, you checked it off, but really how are you practicing that, that’s really the important question.

Clara: As far as the peer pressure, there are eight out of ten teachers who have wicker posters up or are doing word walls. That to me is an arbitrary number that has in no way to do what I’m doing in my classroom. How does that affect me? Me seeing that number, especially with the way learning walks have been done on our campus in the past, it’s very, what do you call it, anonymous…. I think our biggest issue is being able to ask my kids, is this working or you and that honestly, me not being a data person, that is how I’ve figured out for the past however long I’ve been in [the district], I can’t remember right now, how things are working for my kids. I know how things are working for my kids because I see them and I ask them.

While recognizing that such localized reflection may work for teachers’ own practices, Clara, and, throughout the next several months, more members of the design team recognized the need for more systematic ways to monitor implementation. By the time full implementation began in Phase 3, Clara reflected,

For the purposes of this study, I don’t think we can in good conscious say to [the district], this is something that is working for us if we have not figured out a systematic method of checking that the things that we are actually doing, that we are planning right now worked. And if we don’t come up with a method to check did they work, we can’t tell them this worked for us so we should do this too.

Initially, design team members perceived measures of the SOAR innovation to be invalid, as they would not adequately measure the depth of teacher practice they were trying to change.

Further, concerns about perceptions of data misuse in the district continued to shape how they viewed data for improvement of SOAR. Educators struggled to link SOAR with clear outcomes, both as a result of insufficient capacity and their perceptions of the validity of outcome measures.

Although the external partners urged design team members to develop measures and data
collection plans from the middle of Phase 1, the social conditions of data use provided a barrier to using data for their school improvement efforts. With concerns that quantitative measures would be manipulated at higher organizational levels in the district, it was not until late in the first half of Phase 2 that teams began to seriously consider ways that they would measure SOAR.

Using Data for School Improvement

In the previous section, we described how the teacher capacity and the social context of data use in the district each provided barriers to the design team members’ use of data for school improvement efforts. In this section, we describe the transition to using evidence as part of the continuous improvement process. While educators on the development team gradually had a greater openness to data use, teachers on the development teams were sensitive to the ways in which their colleagues would regard their data collection efforts. They were cognizant of the ways in which any data teachers were asked to collect would be seen as part of the accountability-driven district culture. Initially, development team members in non-instructional positions ignored these concerns. They felt that if they shifted their language to emphasize the ways in which the data they collected would be used non-punitively, the staff at their school would be more sympathetic. Frank, an innovation school member in a support staff role at Valley High School, states,

It seems like that’s the language that would make this easier. What we want to see is how it’s being done and how we need to change it. Not, you’re being held accountable whether or not you’re doing it, we want to know how’s it going, what do you need to change.

Other members of the team recognized the need to convey to their staff that they will be collecting and reviewing data related to the SOAR innovation but that it will not be used for accountability purposes. They emphasized how their feedback would be vital for the process of improving the innovation, so it is usable for teachers and leads to positive student outcomes.
Leah, an administrator at an at-large school, summarizes this idea as, “Evaluation of you as a teacher versus the evaluation of the program.” Paulette, an administrator at another at-large school adds, “We’re monitoring to see what works well, what doesn’t work well… If somebody knows I’m doing this to see if this is successful, I want to see where the gaps are, I want to see how to make this better. That’s a different motivation for PDAS [Performance Development and Appraisal System].”

By making this distinction between data for teacher accountability and data for improvement of the innovation, participants were able to revisit the need for some sort of accountability mechanism, even if it was framed in a way that was improvement oriented. The emphasis on having some accountability structures was expressed primarily by the administrators on the design team. Leah continues, “There has to be an accountability piece, otherwise I’m not going to do it because I don’t want to. There has to be some level of, you know, we want you to do this. If you’re struggling, we need to know how to help you but this is a requirement, you must do this.” Paulette explains that they could find ways to monitor implementation through administrators’ routine classroom visits. She elaborates, “The administrator staff looks at how we are going to monitor that, even if it is just a non-punitive checksheet. I saw it implemented this way. The students said x, y, and z and then a quick conversation about it.”

Teachers on the team were skeptical that the development teams could convert the innovation into measurable teacher practices and that teachers in their schools would be receptive of the collection of data about their performance, even if it was to be used to improve a school-developed initiative. Instead of collecting data of teacher practices through classroom observations, they gravitated towards teacher self-reports. When development team members
involved new teachers in piloting the innovation during Phase 2, they relied on perceptual
evidence to inform their improvements of SOAR. In the next section, we describe how teachers
came to rely on perceptual evidence, both to improve SOAR and to build teacher buy-in.

*Types of Evidence for Improvement*

The third research question focuses on the types of evidence educators on the
improvement teams used to make claims about their learning. Despite the emphasis on studying
student outcomes that could be linked to the positive influence of the SOAR innovation, teams
instead gravitated towards perceptual evidence, including formal and informal feedback from
their colleagues. They used this perceptual evidence as a means to refine the innovation design
during Phase 2. Allison, a teacher at Desert Grove, explains, “I think one of the big things, our
first PDSA cycle was very disciplined from what we tried out, and one of the great things was
we got results that we were looking for. I mean, we were hoping that it would be a certain thing.
We were hoping teachers would like it and feel like it was useful and they did.” When design
team members drew on feedback from their colleagues, it helped to customize the design to meet
the needs on that specific campus. At Desert Grove, Jillian shares, “I think—I mean—going
through the [PDSA] cycle we’ve changed the lessons, and—or not changed them, but refined the
lessons to fit what our campus is going to be using for professional development, which then
means we’re changing it to fit our campus.” By eliciting feedback from their colleagues,
development team members could customize the design to meet the needs of their teachers.

This perceptual evidence was also used as a tool to build teacher buy-in. Molly, another
teacher on their team, describes how this feedback was treated as a tool to build buy-in so
teachers “feel like they have a say or like they’re being heard.” At Forest Glen, Rebecca adds,
“[T]here would always be like the data collection part, and then as we were, you know, kind of
looking over the data, that’s when we would like come up with the next PDSA thing, year. We were always collective and looking at data and figuring out what our colleagues wanted, needed, and trying to base our next actions based on that.” At this point in the development process, teams did not consider formal student feedback or look for evidence of change in student outcomes. They focused on drawing on this perceptual evidence to improve the design in ways that better met the needs of their teachers. In other words, PDSA was treated as a tool build teacher buy-in.

The focus on teacher feedback continued throughout the entire development process, as teams felt it was the most beneficial for making improvements to the innovation on their campuses. For instance, at the end of the year members at the Desert Grove and Forest Glen and design teams piloted two separate activities. Forest Glen design team piloted an activity where students filled out a grade reflection form when receiving their progress report. The Desert Grove design team piloted a behavioral reflection designed for teachers to hold students accountable for their behavior, rather than routinely sending them to the front office. In describing their reflection strategy there is a notable absence to student feedback or outcomes. Rebecca shares, “As far as teacher feedback, something they had discussed was how are they going to collect data on something like this, I don’t really want to sort through thousands of notes. So they used some sort of teacher feedback—their perception if the activity will increase ownership, things like that.” Allison says, “We had another meeting a few weeks later where teachers gave us feedback on it. The disciplinary sheet, the feedback we got was that it works really really well on freshman and sophomores.” Although this teacher feedback helped in the alignment of SOAR to the culture of the innovation schools, teams did not examine how the
design may influence student outcomes before spreading it to the whole school in the next school year.

The lack of attention to student outcomes was related to a suspicion of how their initial implementation efforts would affect student outcomes. When discussions of administrative data did arise during the later part of Phase 2, team members remained skeptical. Mark said of data analysis in the context of PDSA in an interview:

I think I would make it a little bit longer process so that we did it in terms of years...right now we’ve been doing it in almost six week or nine week cycles. So at that point it creates a lot of paperwork. And it gives a lot of data that could be viewed either way...Could you just say well, hey, test scores haven’t gone up at the end of six weeks that you did this Plan, Do, Study, Act, so this process isn’t going to work.

Monica adds that the brief PDSA cycles do not map onto the six week grading cycles, wherein most administrative data is collected:

what are we comparing -- if we're looking at change, we're comparing students' change from the beginning of the six weeks to the end of the six weeks or -- Because it seems like comparing this six weeks versus last year's same six weeks is just completely different data, right? Like totally discontinuous. So that's my problem with the PDSA cycle.

Design team members at Valley High School viewed the use of student data more positively, feeling that it too could be a way to garner support for the innovation. Frank says,

What we’re doing right now is knowing, ok, this semester, we’re going to have a few of our design team teachers implement these lessons and the data we collect not only allows us to recognize if this is working but also allows us to say, maybe we can’t prove a direct correlation but the fact that I began using growth mindset and we’ve been talking about it in my room, I’ve seen homework completions go up, I’ve seen tardies go down. It’s a way we can also sell this.

This approach, however, was not put into practice during the development of SOAR, as teams were dependent on perceptual evidence to improve the innovation and build buy-in. At the beginning of school-wide implementation, the design team at Valley used evidence from initial implementation to champion one practice in the innovation, a disciplinary form called the
GROW sheet. Sandra explains, “I think they're going to have to see it to believe it.” Where PDSA fits in. I think the GROW sheets – they're seeing it to believe it. They're believing it because it's proving – it's – it's proving itself. You know, the infractions are down, the kids are actually out there doing something, you know? That type of thing.” As all schools initiated school-wide implementation, this use of student outcome data was introduced alongside perceptual evidence.

When two of the three schools began using student outcome data, it was used alongside perceptual data that team members drew upon to inform their ongoing improvements to the design. The experiences at Desert Grove and Forest Glen demonstrate the continued primacy of perceptual evidence. At Desert Grove, the team relied on a variety of evidence to inform their conversations around a grade monitoring activity where students would routinely track their grades and make plans for how to improve them. Their design team was the only one to make use of data on student outcomes, program implementation, and staff perceptions. For student outcome data, they compared school-wide passing rates during six-week grading periods in the 2013-2014 school year compared to the current year. They monitored implementation by collecting samples of students’ grade monitoring sheets. Perceptual data was gathered through teacher surveys and also asking the staff for suggestions for improvements during faculty meetings. They frame their decision making about how to improve this activity as being informed by each these different sources. Looking at school passing rates, they find only minor changes compared to the prior school year. This lack of expected change in student outcomes and feedback from teachers led the team to make changes to the language on the grade reflection form but also add a more systematic way for students to not only reflect on their grades but make
a clear plan for how they would improve them in the next grading cycle. How to make these improvements was informed by perceptual evidence.

Members of the design team at Forest Glen describe using teacher feedback as a tool for monitoring implementation and better understanding teachers’ perception of the intervention. For a PDSA cycle for a behavioral infraction form—the “Think It Out” form—they surveyed ninth grade teachers to better understand how it was working in this grade. They asked teachers the following questions: “Per week, how many think it out sheets do you sign? Think It Out sheets prevent what percent of your students from needing another intervention? Do you think it would be beneficial to move to a campus wide—do you think other kids, other grade levels would benefit from it?” Their team had collected survey results from students about other practices, but not this form. This brief teacher questionnaire allowed the team to conclude that ninth teachers believed this intervention helped prevent disciplinary issues in the classroom from escalating and that other teachers in the school should use it. For the Forest Glen team, this perceptual evidence was sufficient in determining that the disciplinary intervention should be scaled school-wide, even without evidence that it actually improved classroom behavior or reduced disciplinary infractions.

Yet, later in Phase 3, when presenting to the district design team, Justin apologized for still not having data, by which he meant district administrative data. The team continued to identify data as inclusive of only those things to which they were held accountable. During the Spring 2014 fieldwork visit, one of the leaders of the group, Rebecca, makes reference to the “Think It Out” form in an interview: “As far as the – as far as the Think It Out stuff, that was the first time we had actually had, like, real data from ninth grade and I was like, I was really kind of floored by the reduction as far as, like, number of infractions.” At the same time, their design
team had collected informal teacher feedback in faculty meetings and student survey data, but they did not frame it as data. This point is indicative of tension between data for accountability versus data for improvement observed earlier in the process. Teams came to learn to value and use evidence for improving the innovation, but still felt constrained by the context of the district to share evidence from the accountability system with those outside their team. For instance, Justin went on to reference evidence they had that 80 percent of their students report checking their grades at least once a week. Yet, through conversation with teachers at their school, they came to recognize that student awareness of grades would not lead to higher performance unless it was accompanied by improvement plans and supports from teachers to reach these goals. Design team members drew on different sets of evidence to inform their internal learning (perceptual evidence) and the types of evidence that they shared when they presented their learning to other stakeholders in the district. While development team members came to value perceptual evidence for improving the innovation, they did not think of it as “data” and so, when asked to share findings from their school, they felt as if they had to share outcome data. Teams may also have felt constrained by the context of the district in which outcome data was seen as most valid.

In presentations to district central office staff, they describe the iterative, contextually-based nature of their design process but do not emphasize the improvement data they have drawn upon. Instead, they highlight administrative outcome data. Members of the Desert Grove design team were most explicit, although all teams engaged in it to some extent. When presenting to the central office staff about securing resources to continue their work and scale it to additional schools in the district, they mirrored an earlier presentation made by an administrator in their school who emphasized the positive changes in student outcomes to their peers. An administrator
at Desert Grove, added to the school team’s presentation: “I will share one small bit of data. The first semester our passing rates for our ninth grade students who are traditionally our most difficult students, especially when they are coming in from middle school and they’re trying to get the high school way of things, they’ve increased a minimum of 12 percent across the board” compared to the previous school year. They presented extensive data on the student outcome most aligned with their innovation. They documented improved passing rates in every grade, and nearly every subject. The district’s accountability context is reflected in comments from the other innovation schools. A design team member from Valley explained how at their school, “Behavior of our kids has improved and looking at the data, disciplinary referrals have been down.” An administrator at Valley reiterates the benefits from the use of the behavioral infraction form. She says, “The discipline rate has decreased significantly” and the freshman path to graduate is up to 95 percent. The principal at Forest Glen qualifies why they haven’t seen positive increases in these measures at their school: “So our failure rate is not to say, you know, we corrected in by 15 percent or anything like that. We already had a good passing rate. And we had a good attendance rate. And we had a good discipline rate. It’s not that it’s making that kind of impact, but performance related to core courses, well all of the courses, we are seeing that difference.” When presenting to the district, teams were sensitive to the district’s expectations for accountability as driving higher-level decision-making.

In the context of continuous improvement, school teams distanced their improvements to the design from the PDSA process itself. Within PDSA, there is supposed to be a clear link between the change idea and the outcomes that are being tested. When planning for any change idea, design teams continued to frame their results in terms of student outcomes. Since they were not looking at outcome data to inform improvements to the innovation, improvements were not
necessarily based on data as it was defined in their PDSA plans. Instead, as we’ve discussed, these decisions about how to revise an element of the innovation were instead based on perceptual data from students and teachers. Improvements took place alongside PDSA, rather than as a direct result of the practice. Nevertheless, we found evidence that participants from the innovation schools recognized the value of collecting stakeholder feedback and iterating on the innovation practices. Even without following the specific protocol, teams still continued to identify problem-specific solutions that were user-centered. During an end-of-year activity, Mary asks what they have learned that they could share with the new schools starting this work. One participant said, “because of the PDSA cycles, people are willing to fix problems and find solutions.” Another participant pipes in that “you need to get feedback from both teachers and students to make changes and improvements.”

**Discussion and Conclusion**

Educator evidence use is often portrayed as drawing on a mix of evidence as well as “gut feelings” about what would work best for one’s students. The typical context of this use of evidence is in regards to the instructional routines used in their own classroom. In this paper, we describe how educator teams use evidence for school improvement through the development of an innovation to increase student ownership and responsibility. We describe two barriers to teacher use of evidence: educator capacity to convert broad innovation design into specific outcome measures and the lack of validity of data. Development team members were concerned about measuring SOAR because of unease about measuring any teacher attitudes and practices as well concerns of perceived misuse of data prevalent in the accountability-driven district culture. The district’s social conditions of data use generated a barrier to using data for their school improvement efforts as development team members felt outcomes data would be manipulated at
higher organizational levels in the district, misconstruing local successes (and failures) related to SOAR. Over time, development team members came to appreciate a distinction between data used for accountability and data used for improvement efforts. With this recognition, they began to draw on a range of evidence, while gravitating towards perceptual evidence gathered from teachers. They used this data to make small improvements to the innovation but also to gauge the level of teacher buy-in and make modifications that would better meet teachers’ needs. Outcome data was only drawn upon when trying to prove that the SOAR innovation helped in improving student outcomes. It was not drawn about as a means to improve the innovation.

Previous research suggests numerous barriers to the successful use of educator data use. We add to this discussion a focus of the ways in which the broader social context of evidence use shapes its usage. We find that educators faced numerous challenges in how to use evidence for school improvement efforts, when the district’s data usage tended to be for accountability purposes. Fearing that teachers’ would deem classroom observations evaluative, development team members adopted less intrusive data collection tools, such as teacher surveys. Development team members also found this perceptual data as more informative to their improvement efforts than other forms of data. Though resistance to data use and evidence-driven practices amongst teachers decreased over time, and they even came to value data use to inform and drive improvement efforts, they continued to struggle as they had to straddle the overarching context of accountability, despite their willingness to see data use through the lens of continuous improvement.

The ways in which educators used data raises implications for the ongoing use of the continuous improvement approach to reform. An integral piece of continuous improvement is an understanding of the ways in which local context enable program success. To understand
contextual factors, researchers and educators alike must draw on various forms of evidence. As continuous improvement continues to expand as an approach to reform within schools, it is important for scholars to understand the particular contexts in which practitioners are asked to consider evidence. We find that the context of accountability is ubiquitous and its impact on continuous improvement efforts is one that must be better understood, as it shaped how educators used evidence, often in unintended ways.
Bibliography


<table>
<thead>
<tr>
<th></th>
<th>District high schools</th>
<th>Desert Grove High School</th>
<th>Forest Glen High School</th>
<th>Valley High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student enrollment</td>
<td>20,000</td>
<td>2000</td>
<td>1000</td>
<td>1500</td>
</tr>
<tr>
<td>Student race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>60%</td>
<td>40%</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>Percent African American</td>
<td>20%</td>
<td>20%</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>Percent economically</td>
<td>70%</td>
<td>50%</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>disadvantaged</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent LEP</td>
<td>10%</td>
<td>&lt;5%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>Teacher race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>20%</td>
<td>10%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Percent African American</td>
<td>30%</td>
<td>20%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>Yrs. teaching experience</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

Source. District administrative data, 2012-2013 school year. Pseudonymous school names. All values rounded to preserve the confidentiality of the schools.
<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Phase 1-Design</th>
<th>Phase 2-Development</th>
<th>Phase 3-Implementation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio Files</td>
<td>62 h 22 m</td>
<td>66 h 21 m</td>
<td>61 h 50 m</td>
<td>186 h 33 m</td>
</tr>
<tr>
<td>Field Note Logs</td>
<td>28</td>
<td>24</td>
<td>19</td>
<td>71</td>
</tr>
<tr>
<td>Distributed Artifacts</td>
<td>147</td>
<td>236</td>
<td>74</td>
<td>457</td>
</tr>
<tr>
<td>Research Reflection Forms</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Participant Feedback Forms</td>
<td>97</td>
<td>141</td>
<td>82</td>
<td>320</td>
</tr>
<tr>
<td>Interview Transcripts</td>
<td>24</td>
<td>23</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>Fieldwork Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation team members</td>
<td>21</td>
<td>24</td>
<td>21</td>
<td>66</td>
</tr>
</tbody>
</table>