Planning and Changing
Vol. 51 No. 1/2, 2021, pp. 39–59

MEETING THE EVIDENTIARY NEEDS OF SCHOOL-UNIVERSITY CO-RESEARCHERS IMPLEMENTING THE NEXT GENERATION SCIENCE STANDARDS

The paper highlights the development of a collaborative formative assessment scoring process in a partnership between an urban university and one of the nation's largest districts. We explore collaborative research through the lens of a single formative assessment rubric derived from the Claims, Evidence, and Reasoning Framework (CER, McNeill & Krajcik, 2011) to guide teachers to meet the instructional demands of enhanced learning standards through a consensus scoring process. Results suggest that the formative assessment practices (i.e., using a Next Generation Science Standards (NGSS)-informed science rubric to focus collaboration) reinforced teacher and student learning meaningfully, supporting the enhanced instructional demands of the NGSS and providing school and university partners with useful data for their distinct purposes.

Keywords: School-University Partnerships, Professional Development, Formative Assessment, STEM Education

Introduction

In 2015, the Every Student Succeeds Act (ESSA) was authorized, ending the iteration of the Elementary and Secondary Education Act known as No Child Left Behind (NCLB). This paper highlights the development of collaborative formative assessment as a foundational professional learning process through a single example, a partnership between Loyola University and one of the nation's largest districts, the Chicago Public Schools (CPS). The Loyola-CPS partnership showcases several features of NCLB implementation in Illinois intended to build and sustain capacity for school-level renewal in math, science, or literacy instruction. Through this case, we examine the statewide, multi-tiered systems of assessment and evaluation that we collaboratively developed and applied as a shared evaluation philosophy in which formative assessment by teacher teams was encouraged. Our evaluation philosophy was fundamentally shaped by a belief in the collaborative development, refinement, and enhancement of assessment and evaluation capacity at the program, project, and school levels. Overall, our collaborations enabled state-level meta-evaluators, university partners, educators in schools, and, ultimately, P12 students, to have the evidence they needed to support learning, facilitate systemic improvements, and provide program- and project-level accountability. A chief objective of our efforts was to foster evaluation capacity systemically (Preskill & Boyle, 2008) that could be sustained at the end of

NCLB funding. The 2015 ESSA does not use school-university partnerships as a policy lever for professional learning, so understanding what partnerships accomplished in the NCLB years matters to those who still believe in their promise. It is the multi-tiered approach that drove a statewide system of evaluation capacity building showcased in a single project that we address here and draw some tentative conclusions about what our partnerships accomplished. The overarching challenge was the Holy Grail of professional development evaluation: Can we demonstrate that student learning resulted from teacher learning?

The paper also documents what we learned about school-university collaborative research using common frameworks and tools to meet the evidentiary needs of partners to support and sustain collaboration focused at the school level, highlighting the role of models or frameworks and tools (Leslie, 2011), and protocols (McDonald, et al., 2003) applied in embedded systems of collaboration, described below. Ideally, tools and protocols convene partners in "the day-to-day work of improving teaching and learning" (Bryk, 2009, p. 598). By looking at the partnership through the lens of formative assessment protocols derived from the Claims, Evidence, Reasoning (CER) Framework (McNeill & Krajcik, 2011), we explore how teachers and coaches collectively addressed the enhanced instructional demands of the Next Generation Science Standards (NGSS) and how multiple demands for evidence of learning and improvement were met as a result. Statewide, this and other projects effectively built collaborative research capacity in two ways: 1) convening university staff and teachers as institutional partners to promote sustainable collaboration and 2) using tools and protocols to clarify instructional shifts and make results visible (Hattie & Yates, 2014). Formative assessment protocols and tools impose constraints that support group learning (McDonald, 2003) that are enhanced by ongoing structures and processes focused on instructional enhancement. The formative assessment process used by teachers and university coaches paired a rubric using the CER Framework and NGSS Science and Engineering Practices (SEP) with a consensus scoring process as a research protocol useful for learning how to faithfully implement the NGSS.

The statewide program began as an iteration of NCLB Title IIA professional development block grants to state higher education agencies such as the Illinois Board of Higher Education (IBHE) which developed in three phases of evaluation enhancement from 2004 to 2017. First, we offer an overview of the value of frameworks, tools, and protocols used collaboratively in exemplary professional learning systems. Second, we describe the statewide program that supported 34 partnerships in total but ended with prolonged support to just six to support those efforts that we believed could be sustained as professional learning systems with urban teachers whose work in science education is imperative in any system addressing educational inequities. Third, we characterize the multi-tiered systems of

alignment and accountability that required development of evaluation at several levels of analysis. Finally, we explore the case of elementary and middle grades science teams in several Chicago Public Schools served by an Improving Teacher Quality (ITQ) partnership with Loyola University's Center for Science and Math Education (CSME). The case supports the use of frameworks, tools, and protocols that situate the instructional shifts required by enhanced learning standards like the NGSS in collaborative adult professional learning spaces capable of improving instruction, assessment of learning, and program/project evaluation. We conclude by situating the case in the statewide evaluation system as an example of meeting the evidentiary needs of multiple partners, including the funder.

Professional Learning in Science Education

There is a consensus about professional development (PD) in science, much of which comes from the Eisenhower grants' official reports and evaluations which provide an overview of what exemplary science PD looks like (Garet, Porter, Desimone, Birman, & Yoon, 2001). These exemplary practices do not differ dramatically from the general consensus about PD (Darling- Hammond, et. al., 2009; Hargreaves & Fullan, 2012; Hawley & Valli, 1999; Wei, et. al., 2010). One framework used in Illinois ITQ captures this consensus succinctly as exemplary professional learning includes: a) a content focus; b) active learning; c) coherence; d) duration; and e) collective participation (Desimone, 2009, p. 185). Further investigation into science education highlights the importance of pedagogical content knowledge (Bausmith & Barry, 2011; Magnusson, Krajcik, & Borko, 1999; Shulman, 1987) that allows content-focused teachers to engage students' understanding as a critical feature of science content pedagogy. Exemplary professional learning "incorporate(s) analysis of student conceptual understandings and implications for instruction" (Heller, et al., 2012, p. 333), in formative assessment of student work analyzed for the inevitable science misconceptions and variations in the development of science concepts (Heller et al., 2012). The significance of formative assessment writ large is also well-established (Black & Wiliam, 2001; Hattie & Yates, 2014; Wiliam, 2018), particularly where "formative assessment involves individual and mutual participatory appropriation of learning products" (Ash & Levitt, 2003, p. 23) as when teachers and/or teachers and students analyze products collaboratively and engage one another in ambitious teaching and learning of the sort the NGSS requires. In these cases, teacher and student science learning is assumed to be both socially constructed and cognitively mediated (Ash & Levitt, 2003), requiring collaboration focused on "learning products" with analytical tools that support and sustain ongoing learning, particularly where challenging or counterintuitive concepts drive instruction. But where professional development lacks specific tools to support mentoring, feedback may be in-

sufficient to support teachers' science learning (Zubrowski, 2007). Often the work is supported by a specific framework such as the Five Es (Bybee, 1997) or the CER Framework in the present study and tools like: a) rubrics (Koh, 2011) when the rubrics are instructionally useful and can demonstrate educational impacts (Popham, 1997) and b) collaboration protocols that support professional learning by imposing constraints on conversations to make them more productive (McDonald & Allen, 2017; McDonald, 2003). Calls for specificity in use of professional learning tools include those that: a) designate "a system of tools and socioprofessional routines that foster (ambitious) teaching over time" (Windschitl, et. al., 2012, p. 880); b) limit variations in practice into an accepted instructional core that is socially- mediated and part of organizational culture (City, Elmore, Fiarman, & Teitel, 2009); and c) address science (meta-) cognitions by teachers and students and place student (mis)understandings at the center of instruction and assessment (Heller, et. al., 2012). Tools, like rubrics, protocols, shared academic language, and key frameworks, support teachers to penetrate students' misunderstandings to shape practice and ensure that ambitious science instruction results in enhanced student achievement. Tools support shifts in school-wide and classroom-level discourse that allow science concepts to be developed with co-constructed tools having more influence than imposed ones (Smith & Southerland, 2007), and systems that allow teachers to make their own accommodations to assessment practice are stronger supports for science education reform generally (Towndrow, Tan, Yung, & Cohen, 2010). The present study does this: allows for a framework and tools to support professional learning in science focused on conceptual understanding and integration of key NGSS concepts and practices in use by teachers collaboratively examining student work for sophistication of understanding.

Background

In 2003, the Center for the Study of Education Policy (CSEP) at Illinois State University audited the grant evaluation practices of all grants managed by the Illinois Board of Higher Education (IBHE). We found that only 40% of grants even submitted evaluations, a clear indication of just how ineffective their evaluation system was. Of that 40%, few had used evidence that supported claims about the grants. One outcome of that audit assigned CSEP to apply their audit recommendations for program-wide evaluations to the new NCLB federal block grants. The CSEP team then served as evaluation consultants and meta-evaluators from 2004-2017. The meta- evaluator role was novel and asserted that the IBHE should be intentional about the sustainability of grant achievements by enhancing evaluation capacity as an element of a comprehensive evaluation philosophy, described in more detail below in three phases. This implied a systems approach that embedded a set of evaluation practices at each unit of

analysis: a) the statewide program meta-evaluation that resulted in policy shifts in three-to-four year cycles; b) each partnership between a college or university and schools/school districts; c) the school level where teachers could use formative and summative findings collaboratively to support significant instructional shifts; and 4) ideally, evaluation enhanced each school's ability to inform student learning by engaging students in enhanced assessment. While a philosophy of evaluation was embedded in policy design and requirements, no specific data sources were ever required. Instead, the statewide projects and the meta-evaluation team collaborated to create an evaluation infrastructure for mutual support and accountability. In turn, each project worked with teachers to develop evidentiary sources and the tools to gather that evidence in a dynamic system of evaluation improvement embodied in annual cycles of policy enhancement by the IBHE, described below. Ordinarily this approach resulted in a major policy enhancement every three years that resulted from the collective learning of CSEP meta-evaluators, project directors in every corner of the state, and school-based educators working as partners.

The Illinois Improving Teacher Quality (ITQ) Grant: Three Phases of Evaluation Enhancement

In this section, we describe the 13-year Illinois Improving Teacher Quality (ITQ) State Personnel Development Grant through which the IBHE sponsored 34 school-university professional development partnerships. In that 13 years, the CSEP team served as evaluation consultants and meta-evaluators of the IBHE NCLB grant portfolio, in three major phases extending from 2004 until 2017 when the final ITQ requirements were fulfilled. As previously stated, project level evaluations were never prescribed for the school-university partners. Rather, each project explored its own evidentiary needs and developed capacity to gather and use data about student learning and the effectiveness of each school-level program. In one sense, this was the opportunity to allow projects within a grant-funded program to develop their own assessment and evaluation approaches, given the shifting policy emphasis from high stakes tests to random clinical trials since NCLB. Although NCLB occurred within an increasingly high-stakes-test- oriented policy environment, in 2004 there was yet to be an insistence on a "gold standard" that mandates random clinical trials while relegating more classroom-based, locally developed formative approaches to a lower tier status (Vogt, et al., 2011), despite evidence of their utility to support instructional shifts and collaboration. Overall, we conceived our work as enhancing two key features of ITQ projects: 1) alignment with exemplary professional learning practices and 2) accountability through evaluation and the development of evidentiary sources. Sustainability was the goal for both alignment and accountability mandates.

In Phase I (2004-2006), ITQ grants were widely awarded with few requirements, other than an annual project evaluation and a compact between a school or district and a college or university. This phase resulted in tightened alignment requirements, including key features now widely recognized: professional learning is never a once-and-done but must be job-embedded with opportunities for ongoing collaborative learning at a unit of analysis beyond the individual teacher as a school-wide professional learning system (Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2010; Garet, et. al., 2001). This phase ended in 2006 when CSEP took a much more proactive role, working more directly with partnership projects, traveling the state for school site visits to explore how effectively the projects were serving schools. The school continued to be the unit of analysis given that collaborative teacher learning, exemplary as professional development, was how we envisioned the program's sustainability after NCLB funds evaporated. In the Loyola-CPS case, alignment was a given within a multi-tiered system of district- and school-level supports that leveraged funds from multiple grants and the support of many science educators, but the formative assessments that we describe below were yet to be envisioned.

In Phase II (2006-2010), ITQ meta-evaluators required compliance with exemplary practices in professional development (i.e., increased alignment) and enhancements to project evaluations (i.e., increased accountability) but did so without dictating particulars to encourage local formative assessments in which project partners had a stake. In this phase, many projects were non-renewed if they failed to design for ongoing collaboration, use of student learning evidence from the classroom, and school-level capacity building. A key moment in this phase committed the state to program theory evaluation. Program theory asserts that any program, project, initiative, or intervention has an explicit or tacit theory of action or change. An evaluation is an opportunity to test the theory (Chen, 2015; Weiss, 1997; 2000). This requirement was a watershed moment for ITQ. This provided project directors with an opportunity to reconsider project design to implement a project with a fully developed theory that required attention to school- level arrangements (alignment) as well as ways to gather and analyze evidence of teacher and student learning to test the program theory (accountability). In the case partnership between CMSE and CPS, project designers responded with an increased emphasis on alignment and coaching teachers on site. Loyola designers were also among the first in ITQ to use logic modeling, starting late in Phase I as we were developing this evaluation policy enhancement as an effective tool for laying out the parameters of the program and connecting them to the best evaluatory mechanisms to test the program theory in a cycle of continuous improvement.

In Phase III (2010-2017), enhanced standards (i.e., Common Core State Standards and the Next Generation Science Standards) increased the

policy demands to build evaluation capacity through a program theory approach for planning and guiding evaluation. This proved challenging for many projects, although Loyola embraced the challenge and continued to envision their multi- tiered system interactively as a program theory, demonstrated in annual evaluations. This period saw a winnowing of projects that were not in compliance with the sustainability vision for alignment and/or accountability with only six projects remaining. In 2016, the meta-evaluators assembled a list of ten final deliverables which placed heavy emphasis on sustainability through collaborative formative assessment as the final policy iteration from the IBHE (see Appendix A). Finally, in 2017, projects were shaped by ongoing collaboration between projects and meta- evaluators for the final phases of alignment and accountability. Alignment required matching an Illinois initiative for Professional Learning Communities (PLCs). Accountability enhancements included initiatives focused on formative assessment and action research by teacher teams. Always focused on sustainability, the IBHE asked project directors to use the final funding to ensure that schools had what they needed for continuing alignment and accountability post-ITQ. Ultimately, the meta-evaluation team sought to connect teacher professional learning systems and evidence of student learning. This would be impossible without teachers finding useful tools and protocols to use in professional learning structures like PLCs with appropriate evidence that students learned to standards. In the case of the Loyola-CPS partnership, the NGSS continued to provide an impetus for increased alignment and accountability as these standards require profound instructional shifts. Formative assessments that used the CER Framework and incorporated the Science and Engineering Practices (SEPs) served as the basis of tool and protocol creation and application, explored below.

Loyola University's Center for Science and Math Education (CSME) and Elementary/Middle-Level Science in Chicago

Within the featured project, the characteristics of the final phase of ITQ for alignment and accountability can be showcased by considering any of the final six ITQ projects funded in Phase III. In the case under consideration, the Loyola Center for Science and Mathematics Education (CSME) and Chicago Public School (CPS) partners were already very focused on many of the goals/levers that IBHE had espoused over the years when the *Ten Deliverables* were issued in 2016 (see Appendix A). Overall, these ten, collapsed here to five of particular import to the CSME- CPS partnerships, included key features of high leverage instructional practices: 1) high quality science instruction applying curriculum, instruction, and assessment as the key constituents of content pedagogical knowledge; 2) standards-based alignment of that instruction to include, in this case, Science and Engineering Practices; 3) high quality professional learning

systems focused on each school (alignment); 4) assessment and evaluation well-designed to test the CSME program theory, connecting how teacher learning affected student outcomes (accountability); and 5) leveraging multiple grants using the IBHE philosophy to sustain not only exemplary practice but also to leverage alternative funding. The CSME team of scientists, professional development designers, instructional coaches (all former classroom teachers), and educational researchers/evaluators designed and facilitated professional learning focused on middle grades science teachers originally, but by the final project year was serving science educators from K-8, the elementary/middle-level configuration in most Chicago Public Schools.

As part of their focus on IBHE's meta-evaluation and *Ten Deliverables*, two are highlighted in the present case: 1) #2. *Documentation of a research-based assessment system designed and executed to collect and analyze student learning outcomes at the classroom and school levels* and 2) #5. *Documentation of collaborative formative assessment cycles that strategically reengage students on a daily basis as insights about student learning are used to reengage with specific intentions*. In response, Loyola University Chicago's Center for Science and Mathematics Education (CSME) developed a formative assessment project designed to create and evaluate formative assessment tools and protocols for science teachers in elementary school and middle school, based on the NGSS. This project took place over two years, with Year 1 as a pilot year for rolling out the specific tools and processes that were utilized to generate the data in Year 2 that will be discussed in detail below.

In Year 1, 23 teachers from 11 schools participated in four Professional Learning Community (PLC) sessions over the course of academic year 2015-16. The schools participating in Year 1 had student populations that were predominantly Latiné (> 95% of students) and predominantly low income (> 95 % of students received free/reduced lunch). Across the PLC sessions teachers were introduced to instructional strategies that were aligned with NGSS Science and Engineering Practices (SEPs) 6 and 7. One such strategy was the Claims, Evidence, and Reasoning (CER) Framework (McNeill & Krajcik, 2011) that can be used to help students engage with those SEPs. The CSME team designed a rubric based on the CER framework and developed a process informed by the Bear Assessment System (Sloane & Wilson, 2000) and the Tuning Protocol (MacDonald & Allen, 2017) to help teachers collaborate with each other to look at student work. Based on teacher feedback and evaluator input, both the rubric and the process were refined in Year 2. The design of the rubric is such that scores for Claims, for Evidence, and for Reasoning are assigned separately. This makes the rubric usable across grades K-8; for grades K-5 only the Claims and Evidence scores should be used, since according to the NGSS, the expectation for reasoning is not developmentally appropriate until the 6-8 grade band. All data presented below are from Year 2 (the

2016-17 academic year), utilizing the rubric shown in Appendix B.

In Year 2, the 31 teachers who participated in the project for the duration of the school year were from 12 Chicago Public Schools (CPS); ten serving students in grades K-8, one school serving students in grades K-5, and one school serving students in grades 6-8. Demographics of the participating schools were comprised of primarily low-income, Latiné youth (11 schools) and African-American youth (one school). There were ten teachers from the K-2 grade band, ten teachers from the 3-5 grade band, and 11 teachers from the 6-8 grade band. Schools were selected for participation in the project based on their administrators' willingness to support the project's goals. Schools' prior partnership/participation with CSME ranged from four to ten or more years; therefore, the majority of participating schools had prior exposure to the overall ITQ goal of implementing high quality standards-based instruction. The schools had a strong desire to participate in the project's goals for the 2016-2017 school year. However, only six of the 31 teachers had participated in the Year 1 pilot project.

In Year 2, the rubric was introduced to the teachers during quarterly PLC sessions provided in the 2016-2017 school year. At PLC 1, the teachers were introduced to the rubric, and they engaged with the rubric by scoring student work samples provided by CSME. Teachers first watched a video (https://youtu.be/E4eWYg3jrf8) that was made during Year 1, which showed the teachers using the rubric and additional scoring tools during the process of coming to consensus, in order for them to see how teachers engage in collaborative, evidence-based discussions. The scoring process involved teachers individually scoring the student work samples with the CER rubric and then sharing the scores they assigned using samples with groups of 3-5 teachers at similar grade levels. The teacher groups then discussed the samples and the scores they assigned them, and achieved group consensus on the scoring of the student work samples. After this practice round of applying the rubric, teachers were asked to select appropriate upcoming lessons for their own classes that would be assessed using the rubric.

At PLC 2 (Time 1) each teacher brought four representative samples of their students' work (i.e., samples that represented a range of student performance in the teacher's class). These scores were referred to as Original Scores (see Table 1 for more information). A group of 3-5 teachers then individually scored these work samples using the formative assessment rubric. These scores were referred to as Second Scores. Finally, the group discussed their individual scores to work towards consensus. These scores were referred to as Consensus Scores. In addition to the four samples they brought to PLC 2, teachers were asked to use the rubric to score all of their students' work for the assignment.

At PLC 3, teachers were provided formative feedback strategies that were linked to the rubric and could be used to formatively instruct/re-

engage their students. These strategies included working with students in small groups to design models that accounted for the evidence that they collected during their investigations. Teachers also worked with CSME staff to select an appropriate activity for the next round of scoring.

At PLC 4 (Time 2), teachers scored a second round of student work samples using a similar process as in PLC 2 but based on a different scientific investigation.

Table 1

Data Labels and Descriptions

Label	Description
Original Scores	Scores provided by teachers of their own students' work
Second Scores	Scores provided by teachers of other teachers' students' work
Consensus Scores	Scores provided by teachers of other teachers' students' work – achieved through consensus of 3-5 teachers

For Original Scores, 26 teachers individually scored their own students' (n = 628) work samples using the rubric during Time Point 1. Seven teachers individually scored their own students' (n = 135) work samples using the rubric during Time Point 2 (see Table 2). For Consensus Scores, 24 teachers provided their student work samples (n = 93) during Time Point 1. Twenty-seven teachers provided their student work samples (n = 108) during Time Point 2. These samples were scored first individually then scored collaboratively. Twenty of the teachers from Round 1 also provided student samples in Round 2 (see Table 3).

Table 2

Original Scores

	# of student work samples (# of teachers)
Time 1 (Jan)	628 (26)
Time 2 (June)	135 (7)
Total	763

Table 3

Consensus Scores

	# of student work samples (# of teachers)
Time 1 (Jan)	93 (24)
Time 2 (June)	108 (27)
Total	201

Data Analysis

Evidence of impact of the process on teacher knowledge and skills. One way to examine this is to compare the scores achieved by consensus [Consensus Scores] to the scores achieved by teachers scoring other students' work [Second Scores]. In Year 1, we saw the greatest difference between Consensus Scores and Second Scores were observed at Time Point 2. This could be reflective of the teachers grappling more deeply with the rubric at Time Point 2 that they did at Time Point 1, and the growth of teachers' understanding of what constitutes evidence of their students' grasp of Claim, Evidence, and Reasoning. Conversely, in Year 2 there was little variability between the Consensus Scores and Second Scores at either Time Point. This is not surprising because some of the teachers in Year 1 of the project also participated in Year 2 of the project, thus the group as a whole had a greater familiarity with the rubric. Additionally, all of the participating teachers in Year 2 had access to instructional coaches and during coaching visits, coaches had also helped some teachers become more familiar with the rubric by walking through an example with them, scoring sample work together, and discussing their reasoning. They also helped teachers improve their ability to identify relevant tasks for formative assessment.

In Year 2 there was some variability in the Reasoning Consensus Scores on the rubric in both Time Point 1 and Time Point 2. This is not surprising because the Reasoning dimension of the rubric requires the most cognitive demand, which may lead to varying interpretations of this dimension by teachers (and students). It is also significant to note that throughout this process teachers were permitted to change their Second Scores after discussing their scores with other teachers. This may have influenced the Second Scores and made them less heterogeneous and more similar to the Consensus Scores. Approximately 10-20 percent of original Second Scores were changed post the consensus process. (see Table 4).

Table 4

Difference Between Second Scores and Consensus Scores

Rubric dimension	Mean of second scores (standard deviation)	Number of second scores	Mean of consensus scores	Number of consensus scores
Claim time 1	2.40 (.741)	93	2.41 (.967)	88
Evidence time 1	2.02 (.740)	85	2.00 (.883)	78
Reasoning time 1	1.60 (.746)	47	1.50 (.987)	40
Claim time 2	2.58 (.630)	108	2.61 (.748)	105
Evidence time 2	2.06 (.823)	84	2.02 (.892)	104
Reasoning time 2	1.49 (.919)	64	1.56 (.974)	64

Rubric Scales: Claim, 0 = not evidence, 1 = emerging, 3 = proficient; Evidence and Reasoning, 0 = not evident, 1 = emerging, 2 = intermediate, 3 = proficient.

Evidence of reliability of teachers' individual scores of their students' work. There was no significant difference between teachers' scores of their own students' work and teachers' scores of other students' work. This suggests that teachers' individual scores of their student's work were not influenced by the teachers' bias to overrate or underrate their students' performance [see Table 5].

Table 5

Difference Between Second Scores and Consensus Scores

Rubric dimension	Mean of second scores (standard deviation)	Number of second scores	Mean of consensus scores	Number of consensus scores
Claim time 1	2.28 (.954)	80	2.39 (.741)	80
Evidence time 1	2.12 (.923)	76	2.05 (.752)	76
Reasoning time 1	1.71 (.750)	41	1.62 (.728)	41
Claim time 2	2.70 (.873)	20	2.60 (.718)	20
Evidence time 2	2.33 (.985)	12	2.25 (.905)	12
Reasoning time 2	1.69 (.873)	16	1.61 (.810)	16

Rubric Scales: Claim, 0 = not evidence, 1 = emerging, 3 = proficient; Evidence and Reasoning, 0 = not evident, 1 = emerging, 2 = intermediate, 3 = proficient.

Evidence of impact of the process on student performance. Teacher's individual scores of their own student work [Original Scores] were examined at both Time Points in order to measure student growth. In Year 2, the data collection process included the four student samples used in the consensus process and scores on the assignment from the teachers'

entire class (one whole class data-set per teacher). A paired sample t-test was run between the original scores on the Claim, Evidence, and Reasoning dimensions of the rubric. The pairing was accomplished as follows: Teachers' four student samples were coded and matched for both time points; however, classroom data was not coded. Therefore, although the data was collected from the same classes, it is possible that not all the data are paired samples (e.g., students that leave mid-year or join mid-year). Students performed significantly better at Time Point 2 when compared to Time Point 1 on all dimensions of the rubric [$p<0.01$]. This suggests that students improved in their performance on the CER framework throughout the school year [see Table 6].

Table 6

Change in Student Performance

Rubric dimension	Mean of original scores (standard deviation) point 1	Number of original scores point 1	Mean of original scores (standard deviation) point 1	Number of original scores point 2	$p \leq .05$
Claim	2.10 (1.042)	115	2.45 (.881)	115	.003
Evidence	1.77 (.879)	115	2.07 (.956)	115	.006
Reasoning	1.60 (.746)	47	1.50 (.987)	40	

Rubric Scales: Claim, 0 = not evidence, 1 = emerging, 3 = proficient; Evidence and Reasoning, 0 = not evident, 1 = emerging, 2 = intermediate, 3 = proficient.

Re-engagement and sustainability. Beyond student growth there is some evidence that the formative assessment project showed sustainability and evidence of re-engagement. Re- engagement can be examined from multiple perspectives: *coach-teacher*, *teacher-teacher*, and *teacher-student*.

The very nature of the formative assessment project enhanced teacher use of re- engagement strategies with their students. (Re-engagement, like collaborative formative assessment, is one of the *Ten Deliverables*). Teachers introduced their students to the CER framework in the beginning of the school year, and then re-engaged them in the same process later in the school year. Additionally, the project's coaches provided strategies to help guide teachers in the re-engagement process. One example of formative feedback integrated in the project was in PLC 3 when the coaches had the teachers look at a work sample assessed using the rubric. The teachers were asked to determine whether the task engaged the student in SEPs 6 and 7. Additionally, throughout the PLC sessions coaches incorporated examples of developmentally appropriate responses with regard to the reasoning dimension.

Of the 12 schools, seven included the formative assessment pro-

cess in their vertical team meetings, as evidenced by verbal confirmation from the coaches or from the vertical team meetings agenda notes. For example, one school's vertical team meeting agenda reported a goal of the meeting was to deepen the teachers' understanding of the three dimensions of NGSS. Two of the activities during the meeting included reviewing the CER framework and SEPs 6 and 7 and looking at student work while using the CER rubric. During vertical team meetings at the same school, teachers were asked to reflect on how the CER framework supports students in learning SEPs 6 and 7.

Discussion

When different approaches to supporting pedagogical content knowledge are examined as they have been in the Loyola-CPS partnership, trends emerge that link instruction to professional development activities that are intentional about the particulars of science education and the need for supporting teachers to understand both content in general and students' understandings of key science concepts in particular. In the statewide ITQ project, meta-evaluators, project directors, university-based staff, school-based educators, and eventually students were intended to come together under a regime of enhanced learning standards to employ exemplary practices in professional learning and find ways to link teacher and student learning. In this schema, both alignment of professional learning practices based on content pedagogical knowledge and ongoing, supported collaboration and accountability to all stakeholders, including federal funders, could be addressed.

The IBHE and its consultant/meta-evaluators intended program theory applied flexibly to test professional learning designs (alignment) and evaluation processes and structures that would eventually allow the statewide program and each project to make claims that professional learning arrangements indeed improved student learning outcomes (accountability). This did not mean applying the so-called "gold standard" of causal proof using experimental designs, but rather applying program theory in which clarity in making connections between project activities and a rich set of triangulated evidentiary sources through a design logic allows projects to make credible claims for both teacher and student learning (Weiss, 1997). In this approach, program leaders and project designers can surface tacit assumptions about how the project will work, test them, and answer multiple design questions, "not only the what of program outcomes but also the how and the why" (Weiss, 2000, p. 35). Only in this way can sustainability be ensured as continuous improvement is only possible with evidence that answers core questions in real time and in authentic contexts of practice. In addition, the statewide approach to evaluation encouraged a collaborative, multi-tiered systems of collaboration intended to provide evidence of learning to everyone, from federal funders to stu-

dents who need evidence that engages them to take responsibility for their own learning to high standards like the NGSS.

These features of alignment and accountability are evident in the Loyola CSME-CPS partnership, particularly when we look at the application of program theory and the evidentiary sources available to meet the needs of each of the statewide program's core constituents. In terms of alignment, of the *Ten Deliverables* (Appendix A), the Loyola-CPS science education project was a leader among the final six ITQ projects, fulfilling all the alignment policies the IBHE mandated to reinforce teachers' science pedagogical content knowledge, including the school- based nature of sustainable collaborations and an emphasis on formative assessment to link teacher and student learning in cycles of ongoing improvements. In terms of accountability, the connections drawn between key project features, in this case the CER Framework used as a rubric and a protocol for collaboration linked to intended learning outcomes for teachers and students. These connections are not loose but rather make plain what teachers learned because they had a tool within a strong conceptual frame and were allowed to use it in variety of collaborative learning contexts. In the findings above, teacher learning was documented, and that learning was not superficial. It engaged teachers in really looking at science concepts and how well students understood them. In this way, the value of teacher teams, professional development training on the tool, and the intervention of expert coaches was verified in the results that demonstrate that the tool and processes helped teachers acquire key content pedagogical skills through formative assessment, enough so that they were able to re-engage their students. This is crucial because formative assessment that can speak to science at the level of students learning theory (science) and how to apply it (engineering) because their teachers understand underlying concepts and can see when learning is made visible how to intervene to support students to make meanings from scientific phenomenon and imagine applications as the NGSS envisions.

Education policy at the national level has shifted and become much less open to formative assessments with the features of the Illinois ITQ program and the Loyola CSME-CPS collaborators own designs for formative assessment and evaluation. For one thing, partnerships are no longer encouraged and are, arguably, discouraged with universities having diminished status as partners for professional learning, even though it is difficult to imagine science education advancing without the support of universities. Formative assessment too is discredited in favor of the "gold standard" of experimental design, even though this is very difficult for teachers to do collaboratively in schools, the unit of analysis wherein we believe the possibility of sustainable instructional shifts are most likely to take root. This case study offers some advice for how to work locally in authentic ways with tools and frameworks that engage us all in deeper learning of the sort that real reform of science education will require.

Grassroots sustainability is still possible if we hold to what we know about professional learning (alignment) and evaluation capable of testing our unexamined theories (accountability).

References

Ash, D. & Levitt, K. (2003). Working within the zone of proximal development: Formative assessment as professional development. *Journal of Science Teacher Education*, *14*(1), 23-48.

Bausmith, J. M. & Barry, C. (2011). Revisiting professional learning communities to increase college readiness: The importance of pedagogical content knowledge. *Educational Researcher*, *40*, 175-178.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan*, *90*, 597-600.

Bybee, R. (1997). *Achieving science literacy*. Heinemann.

Chen, H. T. (2015). *Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective* (2nd ed). Sage Publications.

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional rounds in education: A network approach to improving teaching and learning*. Harvard Education Press.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession*. National Staff Development Council.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. [Reports - Research]. *American Educational Research Journal*, *38*(4), 915-945.

Hargreaves, A. & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. Teachers College Press.

Hattie, J. & Yates, G. (2014). *Visible learning and the science of how we learn*. Routledge.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook for policy and practice* (pp. 127–150). Jossey-Bass.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333- 362.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, *22*(3), 255-276, DOI: 10.1080/10476210.2011.593164.

Leslie, D. A. (2011, Spring/Summer). Seeking symmetry in a school-university partnership: University of Chicago and Chicago Public Schools—A collaborative approach to developing models and tools for professional development. *Planning and Changing*, *42*(1/2), 120-154.

Loucks-Horsley, S., Stiles, K., Mundry, S. E., Love, N. B., & Hewson, P. W. (2010). *Designing professional development for teachers of science and mathematics* (3rd ed.). Corwin.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge. In J. Gess-Newsome & N. G. Lederman (Eds). *Examining pedagogical content knowledge* (pp. 95-132). Kluwer Academic.

McDonald, J. P. (Ed.) (2003). *The power of protocols: An educator's guide to better practice*. Teachers College Press.

McDonald, J. P. & Allen, D. (2017). *Tuning protocol*. Coalition for Essential Schools. Retrieved from: http://www.schoolreforminitiative.org/download/tuning-protocol/

McNeill, K. L. & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Pearson.

Popham, W. J. (1997, October). What's wrong--and what's right--with rubrics. *Educational Leadership*, 55, 72-75.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22.

Smith, L. K. & Southerland, S. A. (2007). Reforming practice or modifying reforms?: Elementary teachers' response to the tools of reform. *Journal of Research in Science Teaching*, *44*(3), 396-423.

Towndrow, P. A., Tan, A., Yung, B. H. W., & Cohen, L. (2010). Science teachers' professional development and changes in science practical assessment practices: What are the issues? *Research in Science Education*, *40*, 117-132. DOI: 10.1007/s11165-008-9103-z

Vogt, W. P., Gardner, D., Haeffele, L., & Baker, P. J. (2011). Innovations in program evaluation: Comparative studies as an alternative to RCTs. In M. Williams & W. P. Vogt (Eds.), In *The Sage Handbook of Innovation in Social Science Research Methods* (pp. 293-324). Sage Publications.

Weiss, C. H. (1997, Winter). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, *76*, 41-55.

Weiss, C. H. (2000, Fall). Which links in which theories shall we evaluate? *New Directions for Evaluation*, *87*, 35-45.

Wiliam, D. (2018). *Embedded formative assessment* (2nd ed). Solution Tree.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embed-

ded assessment system. *Applied Measurement in Education*, *13*(2), 181-208. http://www.informaworld.com/smpp/content~content=a783685281~db=all

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878-903.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2010). *Professional learning in the learning profession: A status report on teacher development in the U.S. and abroad* (*Technical Report*). National Staff Development Council.

Zubrowski, B. (2007). An observational and planning tool for professional development in science education. *Journal of Science Teacher Education*, *18*(6), 861-884.

APPENDIX A: Ten Deliverables for the Final ITQ Funding Cycle (2016-17)

(From the 2016-17 Renewal Application,
Illinois Board of Higher Education)

Final 2016-2017 deliverables will include the following, and successful proposals will document with specific evidence how each of these deliverables will be achieved in every partner school. This documentation with evidence requires appropriate analysis and specification of implications and recommendations for each school.

1) Documentation of a school-wide system of continuous improvement that builds capacity to assure cumulative improvements in teacher and student learning that includes evidence of a demonstrable commitment to building or enhancing such a system by committed school principals;

2) Documentation of a research-based assessment system designed and executed to collect and analyze student learning outcomes at the classroom and school levels;

3) Documentation of partners' participation in systematic cycles of planning, doing, and reviewing as they examine all aspects of curriculum, instruction, and assessment that contribute to enhanced student learning;

4) Documentation of multiple iterations of cyclical continuous improvement through assessment, using ITQ tools, indicators, and findings as vehicles of teacher learning understood as essential to boost student learning to achieve enhanced standards at the level of teams and school-wide;

5) Documentation of collaborative formative assessment cycles that strategically reengage students on a daily basis as insights about student learning are used to reengage with specific intentions. Re-engagement then becomes an evidentiary consideration at the team and school levels;

6) Documentation of distributed leadership in a standards-based improvement model mediated principally by teachers in two spheres of continuous improvement: 1) classroom engagements and 2) the collaborative world of selecting, defining, and solving problems with colleagues, coaches, principals and other leaders;

7) Documentation of assessing, planning, and implementing collaborative professional learning systems that include university staff and faculty to meet the specifications of the new RFP;

8) An Executive Summary providing context for the school cases as an overview of the means and mechanisms intended to ensure sustainability and institutionalization;

9) Full descriptions of virtual or other means to continue partnership relationships; and

10) Dissemination of documented local systems of learning and ongoing improvement with developed implications as a host or co-host of a conference or meeting, emphasizing local and regional venues including but not limited to ROEs, university-based regional conferences and workshops, statewide content area venues, and others that allow for other Illinois projects, educators, schools, districts, and universities to benefit.

APPENDIX B: CER Framework Rubric

Example: Students articulate a statement that relates the given phenomenon to a scientific idea, including that the speed of a given object is related to the energy of the object (NGSS Evidence Statement, 4-PS3-1)

	Not Evident (0)	Emerging (1)	Intermediate (2)	Proficient (3)
Articulating the relationship to phenomena (Claim) Students articulate a statement that relates the given phenomenon to a scientific idea.	Does not attempt to make a claim.	Makes an inaccurate and/or incomplete claim. "Some objects have more energy than others." "All objects have the same amount of energy."		Makes an accurate and complete claim. "The faster an object is moving, the more energy it has."
Evidence Students identify and describe the evidence necessary for supporting the claim	Does not describe evidence.	Evidence is described, but it either does not support the claim or is inaccurate. "The gong made sound when the ball hit it." "The gong made no sound when the ball hit it."	Evidence is described and some (but not all) pieces support the claim. "The gong made the loudest sound when it got hit with the fastest ball. The ball bounced off and rolled away."	Every piece of evidence described supports the claim. "In our investigation, we had one fast ball and one slow ball. The gong made a loud sound when it was hit with the fast ball. The gong made a softer sound when it was hit with the slow ball."
Reasoning and Synthesis Students use reasoning to describe why or how their evidence supports their claim.	Does not provide reasoning.	Reasoning does not scientifically or logically support the claim. "The faster ball had less energy." "I know the faster ball had more energy because of baseball."	Reasoning is scientific and logical but is incomplete or does not connect evidence to the claim. "The faster ball hit the gong harder."	Reasoning is scientific and logical and connects all evidence to the claim. "The faster ball made a louder sound because it transferred more energy to the gong. Faster objects have more energy than slower objects."

Dianne Gardner Renn is an associate professor at Illinois State University

Rachel Shefner is an assistant provost at Loyola University, Chicago.

Kelly Holmes is a licensed school psychologist at Deleware Public Schools and adjunct professor at Pepperdine University

Stacy A. Wenzel is an associate research professor at Loyola University, Chicago

Eric Osthoff is an educational consultant.