Developing a Validity Argument for Practical Measures of Student Experience in Project-Based Science Classrooms

William R. Penuel

Katie Van Horne

Jennifer K. Jacobs

Michael Turner

University of Colorado Boulder

With the widespread adoption of new science standards based on *A Framework for K-12 Science Education* (National Research Council, 2012), there is renewed interest in problem- and project-based approaches to teaching. These new standards call on teachers to support students in figuring out core ideas and crosscutting concepts of science by constructing explanations of phenomena and solving problems, using science and engineering practices (Schwarz, Passmore, & Reiser, 2017). Anchoring instruction[1] within problems such as figuring out why antibiotic resistance is increasing and projects such as choosing a tree to plant in a schoolyard to increase its biodiversity provides a practical and potentially engaging way for students to meet these new standards (Penuel & Reiser, 2018).

Realizing the promise of project-based learning to engage and sustain students' interest and build science understanding depends on its enactment in classrooms. Successful projects require students' sustained effort, since they support students' incremental knowledge building typically over multiple weeks (Blumenfeld, Soloway, Marx, Guzdial, & Palincsar, 1991). Past research suggests that student experience can

vary widely by student, task, and teachers (Penuel, Van Horne, Severance, Quigley, & Sumner, 2016; Pitts, 2006). Differences in how teachers respond to and develop students' questioning and reasoning, furthermore, contribute to differences in learning outcomes from project-based learning (Fogleman, McNeill, & Krajcik, 2011; Harris, Phillips, & Penuel, 2012).

Improving student engagement and learning from project-based learning in science requires an approach to research and system of measures that directly supports the improvement of teaching practice. Where past studies have focused on developing more general explanations of variations in teaching and learning for a wide range of classrooms, tools and routines of improvement science promise to help design teams make project-based learning more effective in a wide range of classroom settings (Bryk, Gomez, Grunow, & LeMahieu, 2015). In improvement science, the point of research is not just to learn "what can make things better or worse; it is to develop the know-how necessary to actually make things better" (Bryk, 2015, p. 467). Developing and relying on a system of practical measures—that is, measures that are embedded in the practices of teaching and learning, collected and used frequently, and predict important outcomes—is essential to the enterprise, so that we can learn whether efforts to improve teaching are in fact improvements (Bryk et al., 2015).

In this paper, we present a *validity argument* (Kane, 1992, 2001; Mislevy, 2007) for the design of a system of measures for improving project-based science within a long-term research-practice partnership, the Inquiry Hub. We describe the concrete aims of our partnership for improving science teaching and learning, the system of measures used, and how we are developing evidence to support claims that (1) the measures are both

usable and valuable to teachers, (2) data from practical measures can be collected frequently and used by teachers to improve practice, and (3) the measures predict important outcomes. We center our discussion on validity evidence we are gathering from a student survey within this system of measures that seeks to gather data on students' experience of curricular coherence, relevance, and learning from others in the classroom.

Components of a Validity Argument for a System of Practical Measures

Although the notion of assessment as a form of reasoning from evidence is central to contemporary conceptions of validity, assessment research focuses primarily on developing validity arguments for accountability (e.g., Hill, Kapitula, & Umland, 2011) and research purposes (e.g., Bell et al., 2012; DeBarger et al., 2015). Work to develop validity arguments for measures used to support improvement purposes is just now beginning, and it necessarily draws on both the logic and some techniques used in the past for developing validity evidence. For example, research teams have used cognitive interviews to establish congruence between assessment designers' intent for questions and respondents' cognitive processes when answering questions (e.g., Kochmanski, Henrick, & Cobb, 2015). What is needed is to place these strategies within an overall framework for developing a validity argument—what might be called an "argumentative grammar" (Kelly, 2004) that we might be able to just to evaluate validity arguments for measures that are used for the purpose of guiding improvement efforts.

The general types of claims that need to be supported as part of a validity argument can be inferred from writings on the purposes that practical measures are intended to serve (Bryk et al., 2015; Yeager, Bryk, Muhich, Hausman, & Morales, 2013). One is that

measures must be embedded within the practices of teaching and learning, that is, integrated within the ongoing flow of activities in classrooms, schools, or wherever else learning might take place. Second, the data from practical measures must be focused on specific recurring practices and used frequently to support improvement of those practices. Third, the data must predict important outcomes of the improvement effort. Below, we elaborate on these "claim-types" and the methods and findings we might draw upon to warrant them for practical measures.

Embedded in the Practices of Teaching and Learning

There are many opportunities for embedding measures within the flow of teaching and learning. For example, teachers might report on practices they used in the context of analyzing data on student learning from common assessments used in a teacher collaborative team (Supovitz & Sirinides, 2018). This might be accomplished through the use of an instructional log, where teachers report on practices they implemented or lessons modified and implemented (Krumm & Moorthy, in preparation). Alternately, an instructional coach might use a short protocol to gather data on a specific instructional practice (Bryk et al., 2015). Data on student experience might be gathered through a brief survey (Kosovich, Hulleman, Barron, & Getty, 2015), or an "exit ticket" that a teacher might employ as part of a brief assessment activity (Morozov et al., 2014).

Just because its creators design a practical measure that can be embedded in teaching and learning practice does not mean that it will be successfully. Practical measures need to be *usable* in settings of practice. Usability has multiple dimensions: (1) learnability, that is, how easy it is for people to use it the very first time they encounter it; (2) efficiency, or how quickly tasks can be performed; (3) memorability, that is, how easy it

is to re-engage with after a period of time of not using it; (4) susceptibility to and recovery from errors, and (4) satisfaction with the experience of using it (Nielsen, 2012). Validity evidence related to usability of practical measures might take the form of survey data collected after training related to confidence in ability to use them, log data on how much time it takes to administer the measure, or surveys of user experience that identify challenges to use and how easy they were to overcome, along with the degree to which the experience of administering them was satisfying, rather than frustrating. To date, researchers have reported this type of validity evidence for practical measures mainly for efficiency, in the form of administration time data (Kosovich et al., 2015; Yeager et al., 2013).

In addition to being usable, practical measures need to be *valuable* to users. Ideally, those collecting data are part of a larger shared enterprise in which they have had a say in defining the purposes for using the practical measures. Within a networked improvement community, this might be realized through having opportunities to participate in and give shape to a Plan-Do-Study-Act cycle focused around a specific change to be made to practice (e.g., Penuel et al., in press). Educators may perceive practical measures to be valuable if it provides meaningful data they can use to inform their teaching, and if it helps them make judgments about the efficacy of a strategy they are testing in their classrooms. Perceived value in this instance likely depends on educators' own purposes, however, which may or may not align fully with members at the hub of a networked improvement community. Therefore, data on teachers' purposes for using, modifying, or not using practical measures is needed to help evaluate the perceived value of practical measures.

The concept of *ubiquity* in computing offers yet another goal toward which practical measures might strive. In that context, ubiquity means more than just present everywhere; it also means "vanish into the background" (Weiser, 1991). Our mobile phones today are an example of ubiquitous hardware: they are everywhere, but they have also to large measure faded into the background, become a part of the fabric of everyday life. When computers—or practical measures—are first introduced, the challenge is for users to "make a reliable working infrastructure" (Star, 2010, pp., p. 610) of them, that is, to integrate them into the flow of ongoing work. Ultimately, practical measures should become ubiquitous in both senses outlined above: commonplace within the contexts in which they are used, and also invisible or transparent, that is, faded into the background of the ongoing work of improvement. Observational data of the ease with which people collect practical measures data can provide evidence related to transparency, and data can also be collected related to the degree to which people perceive their use to be part of "routine practice."

Focused on Specific Practices and Used to Support Improvement Cycles

Practical measures are intended to inform improvement efforts, and as such they should target those practices that such efforts seek to improve. In improvement science, practical measures play a key role within Plan-Do-Study-Act cycles organized efforts to address concrete problems of practice through making targeted improvements to practice: they are a key means by which teams can learn quickly whether improvements are having a desired effect (Bryk et al., 2015). They figure in the Plan part of the cycle, when teams select or create measures to assess the efficacy of changes that they make to practices, and in the Study part of the cycle, when teams analyze data from measures to make

judgments about the efficacy of a particular change strategy and consider what refinements might need to be made to it.

But what practices should be analyzed, and how do we specify them? Yeager and colleagues (2013) argue that there are some important criteria that should be met; these, we argue, constitute an important part of the validity argument as well. First, they argue, there must be an empirical warrant for focusing on the practice; the practice should have been shown to cause outcomes of interest. Sometimes, an improvement team must develop such evidence through their joint efforts, but either way, such evidence is crucial, because "Informing improvement places primacy on evidence that directly links to the specific work processes that are the object of change" (Bryk et al., 2015, p. 98). Second, the measure of practice should be precise enough in its specification to be useful for guiding improvements to practice. This requires an identification of a set of "standard work processes" that can be implemented reliably across a wide range of settings (Bryk et al., 2015, p. 90) that can be implemented and measured practically. Third, the practice should be amenable to change, in that the relevant actors on the team have the needed authority to make changes within the time duration of the improvement cycle.

This third requirement leads to an additional needed attribute of practical measures, namely that they can be implemented frequently, or at least as frequently as is necessary to measure change over time and support cycles of improvement work in a meaningful way. Over time, a successful measure will be sensitive to change over time (Yeager et al., 2013) and provide timely enough data to help "teachers, and those mentoring them, decide what to work on next" (Bryk et al., 2015, p. 98). At present, educational systems rarely provide such data to teachers: even when instructional data are available to them,

the data are not provided in a timely enough manner for them to adjust their teaching on a day-to-day basis (e.g., Supovitz & Sirinides, 2018). Educators need more frequently collected data and different kinds of data that can inform improvements to their teaching than interim and end-of-year tests provide (Penuel & Shepard, 2016; Yeager et al., 2013).

It is not sufficient for data on practice to be collected frequently, though, to ensure improvement. A validity argument for practical measures must also include evidence that teachers can and do use them to guide improvement efforts. Yeager and colleagues (2013) provide evidence that use of a practical measure in a networked improvement community to increase rates of success in developmental mathematics in community college helped the network re-focus its priorities on improving belonging uncertainty (Walton & Cohen, 2007, 2011), which was found to predict course success. In addition to re-focusing improvement efforts, practical measures evidence may be used to return to the evidence-base, to identify interventions that are most needed for students with particular profiles (Kosovich et al., 2015; Yeager et al., 2013).

Predict Important Outcomes

One of the most significant ways that practical measures differ from traditional measures is in how the validity of scores is evaluated. For one, a practical measure is judged on its ability to predict important outcomes, whether or not it has high internal consistency reliability (Bryk et al., 2015; Yeager et al., 2013). The focus in prediction is based on both the need for the measure to be feasible to implement in practice—which argues against including many redundant items intended to measure the same underlying construct—and because the goal is to identify early indicators of the success of an improvement strategy, well before long-term outcomes are actually measured. At the

same time, if the improvement work is successful in achieving more reliable, equitable outcomes, then the predictive power of the measure will weaken over time, as teaching practices improve and as variation in practice is reduced (Yeager et al.).

Yeager and colleagues (2013) describe an effort in their community college mathematics network in which they generated strong evidence of the predictive power of a practical measure. Their measure was derived from a small number of items that made up a "productive persistence index." The index was comprised of items about skills and habits for succeeding in college, students' beliefs about their capabilities in mathematics, and their social connections to peers, faculty, and their course of study. They found that students that had a high score on this index on the first day of class were about twice as likely to fail an examination at the end of the term than were students with low scores on the index. The team, moreover, replicated this finding across a total of 30 different institutions, showing the relationship was reliable across those institutions.

The discovery of reliability of relationships across institutions relates to a second difference from traditional measurement in that in judging score validity, practical measurement scores need to be readily interpretable by improvement teams in different settings. Showing that predictive relationships between measures and valued outcomes is just one way to demonstrate interpretability of scores. Kosovich and colleagues (2015) argue that multiple forms of measurement invariance statistics can support a validity argument for score interpretation in improvement science, including demonstrating that the same factor structure for measures exists across groups or time, that factor structures are invariant, that item intercepts are equal across groups or time, and that error variances are equal across groups and over time.

The ability to identify strong predictors depends on some aspects of construct validity that are well-established and appropriate to apply to the design of practical measures. Substantive validity evidence (Messick, 1995) that supports the claim that respondents engage in the expected and appropriate cognitive processes when completing measures is one such aspect and that the language is comprehensible to them (Bryk et al., 2015, p. 101). Such validity evidence may come from cognitive interviews with respondents (Desimone & Le Floch, 2004). When students or teachers are reporting on classroom processes, there should be evidence that descriptions from measures "closely align with the actual work being done by teachers and students" (Bryk et al., p. 94). Content validity evidence (Messick) establishing that the content is relevant and representative of the target domain is also relevant, and may be developed though arguments grounded in theory and evidence related to the core constructs of interest. We turn now to developing the outlines of such an argument for our measures.

Defining Practical Measures of Coherence and Relevance in Phenomenon-Based Science Teaching

Our project is focused on supporting more equitable teaching in science that supports the development of "three-dimensional" proficiency in science, that is, students' ability to apply disciplinary core ideas and crosscutting concepts to explain phenomena and solve problems. We assess that proficiency using a set of multicomponent tasks (National Research Council, 2014) in which students are asked to respond to a set of prompts related to a scenario that presents a phenomenon for them to explain, such as what might have caused changes to average wing length among a population of swallows living in nests attached to a highway overpass. To prepare students to answer these kinds of

questions, the curriculum students encounter presents them with related phenomena (i.e., that require application of the same disciplinary core ideas) in which students figure out core ideas over the course of a unit.

Our practical measures are intended to help us diagnose student experiences of the curriculum as enacted and guide our design team in improving supports to teachers implementing the curriculum. The primary measure is a brief survey, called a Student Electronic Exit Ticket (SEET), which is intended to be used every 5-6 lessons by teachers with students. At present, the survey is administered through a Google Form. A secondary measure is a short observation form used by the secondary science coordinator as part of coaching visits to classroom. Both measures are described in greater detail below in the methods section.

What we hope to learn from these two measures is whether the student experience of the curriculum is *coherent* and *relevant,* because we hypothesize both of these features are important to student performance on our multicomponent assessment tasks. By *coherent,* we mean that students are engaged in science and engineering practices to address questions or solve problems that students—as a class and in partnership with their teacher—have identified and committed to investigate (Reiser, Novak, & McGill, 2017). Design teams create lesson flows that are intended to help students build new ideas systematically and incrementally through their investigations of their questions, and the lessons build toward disciplinary understandings, but the order of lessons reflects students' evolving sense in which these ideas emerged as their questions led to partial explanations, and then to new questions, rather than the order that a disciplinary expert might impose (Penuel & Reiser, 2018). When units are designed to be coherent from the

student point of view and when teachers support students in generalizing from phenomena they study, we hypothesize, they come to recognize when and how science and engineering practices can be used to make sense of new phenomena and problems, as well as what ideas and crosscutting concepts may apply to those problems.

The key teacher practices that we hypothesize support student perceptions of coherence are those of something Reiser, Novak, and McGill (2017) call the *Navigation Routine* (Figure 1). This routine begins with teachers asking students to reflect on what they figured out last time that helped them make progress on questions the class decided they needed to answer, to explain the anchoring phenomenon. Next, teachers ask students to recall what question(s) they decided to pursue next. Students offer initial ideas on how to pursue the question, and the teacher guides students toward an investigation that the teacher has partly planned for the lesson. Students may finish planning and then carry out that investigation. At the conclusion, the teacher facilitates a discussion in which students capture what they figured out that day related to the unit phenomenon and identify question(s) to pursue next. This routine facilitates coherence—along with the carefully designed storyline itself—by putting student sensemaking about the reason for the activity in their hands and facilitating students in making connections between lessons.

With respect to *relevance,* we seek to understand how and when lessons matter to students themselves, matter to the class, and matter to the community. We distinguish among these different forms of relevance, because we hypothesize these may arise from different sources. The ability to connect the day's lesson to a personal interest or related experience may contribute to the judgment that a lesson "matters to me," and this is likely to vary across students. We seek to reduce this variability by selecting anchoring

phenomena that appeal to a wide variety of students, as evidenced by an interest survey we administer. We imagine the day's lesson mattering to the class when the teacher enacts the Navigation Routine and makes use of student questions to help set the direction of the class or engages the students in debate and discussion over the direction for the day. We also hypothesize that a second routine—enacted at the beginning of a unit—the Anchoring Phenomenon Routine (Reiser & Novak, 2017)—supports students in seeing lessons as "mattering to the class." That is because this particular routine involves creating a public record of questions the class agrees is important to answer, to explain the anchoring phenomenon, which is called a Driving Questions Board (DQB). This collective effort is intended to build ownership over the Driving Question and the associated student questions (Weizman, Shwartz, & Fortus, 2008). Last, we hypothesize that student will judge a lesson to "matter to the community" when the teacher makes explicit connections to or engages students in some aspect of the design challenge that links to an endeavor outside the classroom – such as a tree planting initiative (part of our ecosystems unit) or a World Café (a district-organized initiative linked to our genetics unit).

The Current Study

This paper describes a series of studies still in progress to develop a validity argument for two of the practical measures we have devised to support equitable implementation of our curriculum. One of these measures, a brief exit ticket administered on a regular basis, is also to become a major source of data for measuring progress toward a key aim of our current partnership work in science: to make it so that we cannot predict students' experience of coherence or relevance from knowing a student's race,

gender, or home language. In this section, we describe the initial claims we hoped to support, the practical measures, and the data we have gathered or are gathering to develop evidence related to those claims.

The *claims* we wanted to be able to make about the validity of our practical measures are that:

1. Teachers find the exit ticket measure to be *usable* in the classroom on a regular basis.
2. Teachers find the measures *valuable* to their teaching.
3. Teachers implement the exit ticket measure *frequently* enough to provide a reliable measure of growth or change over time.
4. Student responses to questions about their perceptions of coherence and relevance correlate with *specific teaching practices* that can be observed using a second practical measure, a coaching protocol. (not reported here)
5. Student responses to questions about their perceptions of coherence and relevance correlate with *performance on transfer tasks that measure students' three-dimensional science proficiency.*

Practical Measure 1: Student Electronic Exit Tickets

On a weekly basis, students complete brief exit tickets that are embedded within the curriculum materials. These exit tickets are intended to function as a kind of "practical measure" (Yeager et al., 2013) of student learning and experience—that is, a measure that is feasible to implement within the flow of instruction and that can inform the teaching and learning progress. The intent is for teachers to be able to use these to assess the degree to which students were able to add to their understanding of the phenomenon at

hand what was intended in the written lesson, as well as to measure student experience of the curriculum.

Students report on different aspects of their experience in the exit tickets. They report on its perceived coherence, including their clarity about why they are engaged in the day's lesson and how it contributes to them answering the unit's overall driving question. In addition, they report on the perceived relevance of the day's lesson, indicating whether it mattered to them, to the class, and to the community. Third, students report on their affective response to the lesson, that is, whether they felt bored, excited, confused, or confident during the lesson. Our research to date indicates that the coherence of students' self-reported learning experiences are associated with feelings of excitement and perceptions of relevance to their lives and their community, two important dimensions of student engagement (Penuel et al., 2016). We view the exit tickets as a key component of the assessment system, because it provides evidence as to students' disciplinary learning and their perceptions of how connected the lessons are to one another and to their own lives.

Practical Measure 2: Coaching Protocol

The coaching protocol focuses on the degree to which teachers adhere to the basic steps of the teaching routines around which the curriculum is organized, as well as on students' understanding of the purpose of that day's lesson. It is intended to be used by an instructional coach; for the past year since we began using this measure, the user has been the district secondary science coordinator, who has approval from our institutional review board to serve as a data collector for supporting our curriculum implementation research. We have begun—as part of this series of studies to develop validity evidence for the

SEET—to use this protocol as a more formal data collection protocol, though its purpose at present is to determine whether we can identify any links between teacher practices marked on the coaching protocol and student responses to the SEET.

The coaching protocol is intended to be used as part of a coaching routine that begins with coach and teacher deciding on a focus for the observation, then the coach conducts the observation, and finally the coach and teacher debrief on the observation. Each of the major questions follows the steps of the Navigation Routine and asks the coach to mark whether the step was followed and to write a sentence about how the step was enacted, if it was enacted. Each section has a set of potential feedback-oriented questions that the coach can ask in the debriefing session. The other major part of the protocol that we use to elicit student experience entails the coach asking five different students what they are doing today and why, how the day's lesson relates to the anchoring phenomenon, or—if appropriate given the focus of the lesson—how it helps with the design challenge.

Sources of Data for Developing Validity Evidence for the SEET

We focus on teachers' perceptions and uses of practical measures associated with our evolution unit, unless otherwise specified below. There are both informal (e.g., discussions with teachers during professional development sessions) and formal (e.g., tests that have been reliably scored by expert raters) sources of evidence that inform the argument that we develop here.

Informal discussions with teachers. In fall and winter 2017, as part of a videoconference discussion and one in face-to-face professional development with teachers piloting our curriculum, we facilitated discussions with teachers focused on improving the value of the SEET for their own practice. In both discussions, we solicited

input on changes they would like to see made to the SEET and information on what they perceived to be obstacles to their use. The outcomes of those discussions are presented here; transcripts of these discussions are not available.

Logs of use. Because the SEET is a Google Form to which we have access to student responses generated by teachers, we have records of the frequency with which teachers have used the SEET. The data we focus on for this analysis come from the 2016-17 school year and focus on our evolution unit.

Teacher use and perceptions survey. In spring 2018, we administered a survey to 19 teachers who have been part of the pilot testing of our year-long biology curriculum. The survey is ongoing; so far, a total of N teachers have completed the survey. The survey focuses on teachers' perception of the value of a wide range of assessment tools we have co-created with teachers—including the SEET—for different purposes: for assessing student understanding, for assessing student engagement, and for improving their teaching. It also asks them to self-report on the frequency with which they used different assessments, including the SEET.

Observations using the coaching protocol. On days when the SEET is being used, researchers are using the coaching protocol to identify teaching practices that may be associated with particular patterns of student responses to the SEET. This data collection is just underway and is not part of the analysis presented in this paper.

Transfer tasks. A key learning goal is to support students' generalizing from their investigations of phenomena. It is not enough that students be able to explain a particular phenomenon; it is critical that they be able to abstract science ideas that they can apply to the study of related phenomena. Typically multiple cases are necessary, in order to

facilitate reasoning from cases to develop generalized ideas (Kolodner, 1993; Kolodner, Gray, & Fasse, 2003). For this reason, our units present more than one phenomenon to students, and students are supported in applying ideas from one to a subsequent phenomenon and elaborate on those ideas. In addition, we have developed a set of summative assessment tasks that teachers can use that present students with an unfamiliar phenomenon in which they must use science and engineering practices to explain, applying focal core ideas and crosscutting concepts.

The design of the transfer tasks follow guidance regarding the construction of multi-component tasks organized around a scenario that presents a phenomenon or problem to students (Achieve, 2018; National Research Council, 2014). As an example, for the evolution unit, one transfer task presents evidence of change within a population of swallows that adapted to life beneath a new interstate highway. Students analyzed data about wing length, nests, and road kill to build an evolutionary explanation for how the population might have changed. The students apply what they have learned from two related phenomena involving bacteria and a different species of birds, to answer these questions.

These transfer tasks are integrated into optional unit tests that are provided to schools by the district central office, for the purposes of external monitoring. Through our partnership, we have been able to integrate these tasks into those assessments through the typical review process used to develop these assessments. The district assessments include other tasks, however, that have been developed by teachers not using the units and with more limited preparation in the design of three-dimensional assessments. To support more coherent assessment development, members of the partnership team have

developed professional development supports for district teachers writing assessments, which we are currently studying.

Pre-and post-test data were collected for the purpose of developing validity evidence for the SEET from four different teachers' classrooms who also provided SEET data for one session early in the teaching of the evolution unit. A total of 128 students completed two of four assessment tasks provided to them at pretest and two different tasks at posttest. For the present analysis, we examined correlations between SEET questions related to coherence and relevance and overall scores on the performance tasks.

Results

Usability of the SEET

There are several design criteria that are intended to enhance the usability of the SEET that we have applied and refined since we began using it in fall 2015. First, we have sought to limit the number of questions on the SEET, to make it easy to administer quickly. In addition, we have limited the number of open-ended questions to one per SEET, also to increase efficiency. Second, we rely on the "Google Ecosystem" to administer the SEET, to facilitate integration with other district technology services. The use of Google Forms to collect data enables teachers to modify surveys as they see the need, and it also allows them to capture data from their class in a spreadsheet. To facilitate ease of interpretation of responses, we use primarily yes-no-unsure formats for all SEET questions. This approach allows teachers to read the pie charts that Google Forms generates automatically relatively easily, because these charts show the distribution of responses to each question within their class relatively easily.

One of the factors that limits the usability of the SEET is access to adequate technology in the classroom. Our use model calls for one computer for every three students, not just for the SEET but also for engaging with the computer simulations embedded in the curriculum. Yet in our discussions in professional development, some teachers told us that they did not have this level of access to computers where students could complete the SEET. Having students access the SEET on phones was problematic to some teachers, they said, because of classroom and school policies that limit the use of phones in classrooms.

Another factor that affects how usable teachers perceive the SEET to be is their own comfort with using Google Forms in the way that the SEET makes use of this technology. We have sought to reduce barriers to use by setting up folders for teachers and providing instructions as part of professional development as to how to access the SEET surveys and data, and we have also provided an opportunity for teachers to analyze mocked up data and modeled possible uses of the data. Even so, some teachers find that it takes at least 15 minutes to set up computers and collect data from students using the SEET, far more than the amount of time the SEET is intended to take (between 3 and 5 minutes to total class time).

We thus concluded that at best, we have mixed support for the usability of the SEET at present.

Perceived Value of the SEET

To increase the likelihood that teachers will see value in the SEET for their own teaching, we re-design the instrument each year with teachers who participate in the design process. Over the years, we have changed the content of the SEET to reflect

teachers' desired focus. Last year, for example, teachers have become more interested in more questions related to student experience of collaborative learning in small groups, and so we added questions related to that topic.

This year, we noticed that fewer teachers were using the SEET and so initiated a pair of conversations related to its content, with the intent of increasing its perceived value to teachers. Some said that it needed to be shortened and could be, if we did not include the same questions on every form. Teachers felt asking demographic information of every student each time was especially redundant. We did so only to allow teachers to use the same URL and Google Form for the SEET every time, but this tradeoff in usability and value resulted in less rather than greater use of the SEET by teachers. In addition, the single Google Form meant that those teachers who did want to compare data from different sessions had to download spreadsheet and create representations of the data by hand.

When asked what improvement we could make that would increase the SEET's value to them, teachers suggested that having some questions that they could use for grading purposes would help. In the past, we had heard teachers say they needed more items that they could use could assess student learning, and not just the extended response questions that were common on the assessments embedded within the unit. Implied in their suggestions was the claim that—relative to assessments of science content—student engagement data was less valuable to them as a group.

On the basis of this feedback, we made two key modifications to the SEET. First, we created lesson-specific versions of the SEET, each of which had three multiple-choice questions related to the content of the day's lessons. With those lesson-specific versions,

we were also able to customize each SEET, so that there were a pool of common items (to track change over time) but also so that students saw some different questions each time. In addition, we eliminated the collection of demographic information from all but the first SEET of the unit. As these changes have just been implemented for this year's evolution unit, we have yet to know whether these changes will make a difference to teachers' perceptions of the value of the SEET.

Frequency of SEET Use

We find wide variability in the frequency with which teachers have collected SEET data, in which few teachers use the SEET as frequently as is intended (Table 1). This fact is concerning to us for a variety of reasons. For one, to be useful for guiding our improvement effort, the SEET needs to be used frequently enough to measure whether changes made to practice result in a measurable improvement in student experience. If teachers use the SEET only once per semester or not at all, it is not possible to estimate change in student experience at all among those teachers. If those teachers are in some way different from those that do gather SEET data from students in ways that correlate with valued outcomes, this fact could diminish the efficacy of efforts to identify powerful predictors within the data that are associated with variations in outcomes. Another reason we are concerned about limited use is that makes it difficult for us to be able to detect whether variations in student experience are linked to different teaching practices, because there are too few teachers using the SEET to generate variation along multiple possible dimensions of teaching (e.g., use of particular steps in the Navigation Routine).

Table 1.

Frequency of SEET Administration

Frequency of SEET Administration*	Number of Pilot Teachers
No occasions	8
1-2 occasions	2
3-4 occasions	6
5-10 occasions	5

*Teachers were asked to use weekly with students, up to 10 weeks.

Identification of Powerful Predictors within the SEET

Building on prior research we conducted using only SEET data from an earlier wave of data collection (Penuel et al., 2016), we first analyzed correlations between items that relate to coherence versus relevance. We did find a modest but statistically significant correlation between coherence and relevance items. When students said that they understood why they were doing what they were doing that day, they were more likely to report that the lesson was personally relevant to them ($r = .25$, $p < .05$) and that they were excited by the lesson ($r = .22$, $p < .05$).

We did find significant correlations between student scores on the transfer tasks and selected questions related to coherence from the SEET. As Table 2 below shows, there was a small but statistically significant correlation between scores on the transfer task and responses to one of the three statements on the SEET related to coherence could endorse: "I know why we did what we did in class today."

Table 2.

Correlations of Coherence Items with Transfer Task Performance and Gains

	Relationship to Posttest Scores on Transfer Tasks	Relationship to Gains on Transfer Tasks

"I figured out something today that helped us make progress on the questions on the Driving Question Board."	.113	.107
"I know why we did what we did in class today."	.228*	.177*
"I know what questions we will need to investigate next."	.115	.065

N = 127-128 * p < .05

We found only one significant correlation—also small—when we examined relevance questions on the SEET in relation to transfer task scores. There was a small, positive correlation between students endorsing that the day's lesson "matters to me" and their scores on the posttest. No other correlations were significant.

Table 3.

Correlations of Relevance Items with Transfer Task Performance and Gains

	Relationship to Posttest Scores on Transfer Tasks	Relationship to Gains on Transfer Tasks
"Today's lesson matters to me."	.206*	.155
"Today's lesson matters to the class."	.166	.108
"Today's lesson matters to the community."	.079	.066

N = 127-128 * p < .05

Discussion and Conclusion

We found mixed support for the claims that we have investigated so far in our research to develop a validity argument for practical measures that can support improvement in teaching science in ways that align with the vision for teaching and learning outlined in *A Framework for K-12 Science Education.* Despite being designed

for efficiency and ease of use within a system we assumed teachers would be familiar, teachers reported difficulty using the SEET on a regular basis. In addition, because our original SEET did not include questions that they might expect to include on an exit ticket—questions to elicit students' understanding of the day's science content—teachers found it to be of limited value. And while some teachers used the SEET frequently enough to be able to document changes to teaching, many did not. We were successful, however, in identifying some questions that could serve as significant—if not exactly powerful—predictors of student learning outcomes that were consistent with our hypotheses regarding how coherence and relevance might matter for student outcomes.

In this paper, we have not presented validity evidence related to the coaching protocol, nor have we reported on any analyses indicating whether we can reliably link differences in student experiences to implementation of particular teaching practices. Missing also are data presenting evidence that students respond to the SEET questions in ways that are comprehensible to both them and us in ways that are consistent with our expectations about the ways students' reports of their experience might vary. The lack of such data does cast doubt on the trustworthiness of some of the analyses we have presented here, particularly those that relate to the search for powerful predictors within the SEET. We can only point to some success in replicating earlier within SEET analyses as reasons why the data might be trustworthy. Both evidence related to the coaching protocol and student interpretation of items are now being collected with funding from a new grant from the Spencer Foundation.

With this new grant we will also be developing a digital infrastructure for customizing the SEET, collecting data, and representing data back to teachers. We are at

the very beginning stages of this work, which will employ a user-centered approach to design of our digital infrastructure consistent with our participatory design approach used to organize joint work in the partnership overall.

Finally, we do plan in the fall to identify a professional learning community in a school that is interested and willing to participate in a design study that is focused on supporting the *use* of the SEET data to improve the quality of student experiences in the classroom. Ultimately, even if we are to show that the SEET is usable and has some value to teachers for gathering data they can use to assess student understanding, we will not have succeeded in our effort, unless we are also able to show that teachers can and do *use* evidence from the SEET in ways that inform and meaningfully guide efforts to improve teaching.

Implications for Other Efforts to Validate Practical Measures

The set of studies whose basic findings we have presented here underscore a well-known fact in assessment research, that developing validity evidence for using a new measure for any purpose is painstaking and includes many disappointing results. This ought not dissuade us or others from undertaking such research, because the trustworthiness of our claims about our larger endeavor that we can improve the quality of student experience for all students depends on our having good measures of outcomes. Just because measures are intended to be "practical" does not mean their validation will be as easy to accomplish. Planning a series of studies of measures to coincide with development efforts may be the only feasible way to undertake validity studies of practical measures, however, because of the pace with which improvement research is expected to unfold.

Achieve. (2018). Criteria for procuring and evaluating high-quality and aligned summative science assessments. Washington, DC: Author.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-87.

Blumenfeld, P., Soloway, E., Marx, R. W., Guzdial, M., & Palincsar, A. S. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26*(3&4), 369-398.

Bryk, A. S. (2015). Accelerating how we learn to improve. *Educational Researcher, 44*(9), 467-477.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard University Press.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2015). Building an assessment argument to design and use NGSS assessments to evaluate the efficacy of curriculum interventions. *American Journal of Evaluation, 37*(2), 174-192.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1-22.

Fogleman, J., McNeill, K. L., & Krajcik, J. (2011). Examining the effect of teachers' adaptations of a middle school science inquiry-oriented curriculum unit on student learning. *Journal of Research in Science Teaching, 48*(2), 149-169.

Harris, C. J., Phillips, R. S., & Penuel, W. R. (2012). Examining teachers' instructional moves aimed at developing students' ideas and questions in learner-centered science classrooms. *Journal of Science Teacher Education, 23*(7), 768-788. Retrieved from http://dx.doi.org/10.1007/s10972-011-9237-0

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kelly, A. E. (2004). Design research in education: Yes, but is it methodological? *The Journal of the Learning Sciences, 13*(1), 113-128.

Kochmanski, N. M., Henrick, E. C., & Cobb, P. A. (2015). *On the development of content-specific practical measures assessing aspects of instruction associated with student learning*. Paper presented at the Using Continuous Improvement to Integrating Design, Implementation, and Scale Up, Nashville, TN.

Kolodner, J. L. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kolodner, J. L., Gray, J. T., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in Learning by Design classrooms. *Cognitive Science Quarterly, 3*(2), 119-170.

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *Journal of Early Adolescence, 35*(5-6), 790-816.

Krumm, A. E., & Moorthy, S. (in preparation). The multiple roles of data in research-practice partnerships.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Morozov, A., Herrenkohl, L., Shutt, K., Thummaphan, P., Vye, N., Abbott, R. D., & Scalone, G. (2014). Emotional engagement in agentive science environments. In J. L. Polman, E. Kyza, K. O'Neill, & I. Tabak (Eds.), *Proceedings of the 11th International Conference of the Learning Sciences* (pp. 1152-1156). Boulder, CO: International Society of the Learning Sciences.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Research Council.

National Research Council. (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

Nielsen, J. S. (2012). Usability 101: Introduction to usability. Retrieved from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Penuel, W. R., Bell, P., Neill, T., Shaw, S., Hopkins, M., & Farrell, C. C. (in press). Building a Networked Improvement Community to promote equitable, coherent systems of science education. *AASA Journal of Scholarship and Practice*.

Penuel, W. R., & Reiser, B. J. (2018). Designing NGSS-aligned curriculum materials. Paper prepared for the Committee to Revise America's Lab Report. Washington, DC: National Academies of Science and Medicine.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (pp. 787-851). Washington, DC: AERA.

Penuel, W. R., Van Horne, K., Severance, S., Quigley, D., & Sumner, T. (2016). Students' responses to curricular activities as indicator of coherence in project-based science. In C.-K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Proceedings of the 12th International Conference of the Learning Sciences* (Vol. 2, pp. 855-858). Singapore: International Society of the Learning Sciences.

Pitts, V. M. (2006). *Do students buy in? A study of goal and role adoption by students in project-based curricula.* (doctoral dissertation), Northwestern University.

Reiser, B. J., & Novak, M. (2017). *Developing coherent storylines of NGSS lessons*. Paper presented at the NSTA Area Conference, Milwaukee, WI.

Reiser, B. J., Novak, M., & McGill, T. A. W. (2017). *Coherence from the students' perspective: Why the vision of the Framework for K-12 Science Education requires more than simply "combining three dimensions of science learning*.

Paper presented at the Board on Science Education Workshop “Instructional Materials for the Next Generation Science Standards, Washington, DC.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). Moving beyond 'knowing about' science to making sense of the world. In C. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 3-21). Washington, DC: NSTA.

Star, S. L. (2010). This is not a boundary object: Reflections on the origin of a concept. *Science, Technology, and Human Values, 35*, 601-617.

Supovitz, J. A., & Sirinides, P. (2018). The linking study: An experiment to strengthen teachers' engagement with data on teaching and learning. *American Journal of Education, 124*(2).

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality & Social Psychology, 92*, 82-86.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447-1551.

Weiser, M. (1991). The computer for the 21st Century. *Mobile Computing and Communications Review, 3*(3), 3-11.

Weizman, A., Shwartz, Y., & Fortus, D. (2008). The driving question board: a visual organizer for project-based science. *The Science Teacher, 75*(8), 33-37.

Yeager, D., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). Practical measurement. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

[1] The term “anchored instruction” in reference to problem-based learning can be attributed to the work of the Cognition and Technology Group at Vanderbilt (Bell et al., 2012; DeBarger, Penuel, Harris, & Kennedy, 2015), which developed the idea in the context of mathematics teaching.