Chapter 15
Measuring and Facilitating Highly Effective Inquiry-Based Teaching and Learning in Science Classrooms

Jeff C. Marshall
Clemson University, USA

ABSTRACT

For the last decade or so there has been a huge push to incorporate best practice into the classroom. For science, this includes bringing effective inquiry-based instruction into all classrooms as a means to engage the learner. However, all inquiry instruction is not equal in terms of improving student achievement and conceptual development. This chapter explores how four critical constructs to learning (curriculum, instruction, discourse, and assessment) can be effectively measured and then used to guide more effective instructional practice. The Electronic Quality of Inquiry Protocol (EQUIP) is an instrument that can be used to measure and then to frame the discussion regarding the quality of inquiry-based instructional practice. Specifically, this chapter provides an overview of EQUIP, details the reliability and validity of EQUIP, shares a sample lesson that is analyzed using EQUIP, explores ways that EQUIP can help with teacher transformation relative to inquiry instruction, and addresses the relationship of EQUIP scores and student achievement data. There is a very high correlation between teacher performance on EQUIP and the ensuing student growth noted during an academic year.

DOI: 10.4018/978-1-4666-2809-0.ch015

MEASURING AND FACILITATING HIGHLY EFFECTIVE INQUIRY-BASED TEACHING AND LEARNING IN SCIENCE CLASSROOMS

Bringing high-quality inquiry-based instructional practice into science classrooms has continued to be central to reform efforts for the last several decades (Bransford, Brown, & Cocking, 2000; National Academy of Sciences, 2007; National Research Council, 1996, 2000, 2012; National Science Teachers Association, 1998). However, merely increasing the quantity of self-reported inquiry instruction is insufficient (Marshall, Horton, Igo, & Switzer, 2009); the quality of inquiry instructional practice must be at such a level that teachers are effective in facilitating rigorous, standards-based, inquiry-based learning. Success in achieving this goal has been largely inconsistent at best in programs across the country. Definitions of inquiry-based instruction may vary somewhat, but clear direction has been given to defining and exemplifying inquiry-based instruction (NRC, 1996, 2000). Even though consistency can be found among many of the definitions and agreement is found in the desire for reform that includes inquiry-based instruction, the implementation remains inconsistent.

As science education looks ahead to the next decade or so of science instruction, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012) has begun to pave the way for a new vision of teaching and learning that is more intentional and more integrated. This framework which serves as the predecessor for the *Next Generation Science Standards* (due out soon) makes clear that inquiry forms of instruction need to integrate cross disciplinary concepts and core ideas into the learning. No longer is it sufficient to teach inquiry as a stand-alone unit and then proceed to learning "the content." Inquiry helps provide the context for learning major concepts and ideas.

We know that successful inquiry-based instruction is often the result of numerous professional development experiences. For more experienced teachers, these experiences are necessary because transformation of practice is needed to move from prior more teacher-centered paradigms to a more student-centered, constructivist approach where students build on prior knowledge through a series of science learning experiences. For neophyte teachers, inquiry instruction should not be a new concept, but support is often needed to help achieve a significantly inquiry-centered environment. For these beginning teachers, they often have to overcome many of their college experiences, which were typically solely confirmatory experiences as well as overcoming peer, departmental, and/or school structures that may model learning that is counter to inquiry.

Capps, Crawford, and Constas (2012) conducted an analysis of professional development programs in an effort to see how well aligned to best practices many of the programs currently are. Their findings suggest that most of the analyzed programs generally align with the recommended features of effective professional development (Darling-Hammond & McLaughlin, 1995; Garet, Porter, Desimone, Birman, & Yoon, 2001; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2010). The structural features of effective professional development experiences include significant interactions that include extended support through the academic year while providing authentic experiences for teachers. Further, core features of such professional development include many or all of the following: coherence, lesson development, modeled inquiry experiences, reflection, transference of new skills, and content knowledge development. Our work has shown during the past 5 years, that teachers who are engaged in our professional development model associated with Inquiry in Motion are able to raise student achievement higher than the virtual comparison group of similar students (Marshall, Horton, & Edmondson, 2007; Marshall & Horton, 2009).

These changes in teacher practice bring about the most significant change in student achievement after two or more years. During the first year, teaching practice is often inconsistent and unstable as new instructional approaches are used, and the data suggest that during year two that as teachers become more consistent in implementing guided forms of inquiry-based instruction that learning goes up significantly.

In helping to facilitate highly successful inquiry-based teaching and learning, two components are essential: 1) a measure is needed that provides an understanding of the degree of success experienced and tracks any progress made, and 2) a plan is needed for how to successfully implement inquiry-based instruction. Regarding the metrics, numerous instruments exist that monitor and measure teacher effectiveness and include the following:

- *Inside the Classroom Observational Protocol* (Horizon Research, 2002)
    - **Target:** Provides a global view of effective classroom practice;
    - **Possible limitation:** Very long and complex measure—lacks targeting specific aspects of instruction.
- *Reformed Teaching Observation Protocol (RTOP)* (Sawada, et al., 2000)
    - **Target:** Provides measure of constructivist practice in math and science classrooms;
    - **Possible limitations:** The non-descript Likert scale makes interpreting the value of individual and holistic scores difficult, indicators within constructs do not hold together for factor analysis, and entire lesson sequence needs to be observed for accurate score to be determined.
- *Electronic Quality of Inquiry Protocol (EQUIP)* (Marshall, Horton, & White, 2009; Marshall & Horton, 2009; Marshall, Smart, & Horton, 2010)
    - **Target:** Provides a measure of inquiry-based instruction in math and science through the use of a descriptive rubric;
    - **Possible limitation:** Targets specific type of instructional practice.
- *Science Teacher Inquiry Rubric (STIR)* (Beerer & Bodzin, 2003)
    - **Target:** Inquiry instruction aligned to the *National Science Education Standards (NSES)*;
    - **Possible limitations:** All components of inquiry defined by NSES are not present in every lesson, and instrument will be quickly antiquated because it is targets a narrow (NSES) view of inquiry.
- *Science Management Observation Protocol (SMOP)* (Sampson, 2004)
    - **Target:** Measures classroom management issues that support effective science instruction;
    - **Possible limitations:** Classroom management is necessary but not solely sufficient to predict successful instructional practice; so, while informative, this instrument seems to need to be coupled with another instrument for solid understanding of successful instruction to be understood.
- *Secondary Science Teaching Analysis Matrix (STAM)* (Adams & Krockover, 1999)
    - **Target:** Move novice teachers toward more student-centered instructional strategies;
- *Expert Science Teaching Educational Evaluation Model (ESTEEM)* (Burry-Stock & Oxford, 1994)
    - **Target:** Excellence in science teaching from a constructivist framework; and
- **Teacher efficacy scales (Riggs & Enochs, 1990)—target:** Predict likelihood of reform.

Despite all the available metrics, initial research showed that two instruments, the RTOP and the EQUIP, provided solid measures for assessing the degree to which inquiry-based instruction is being facilitated in K-12 science classrooms. These two instruments stand out for several reasons: (1) both of these instruments provide measures that target inquiry-based instruction in math and science classrooms, (2) both have validity studies to support their work, and (3) the other instruments seem too general (e.g., consider all elements of effective practice), too granular (e.g., consider one aspect of instruction such as classroom management), or too complex (e.g., necessary to use multiple rubrics over multiple days). Further, by attending any science education research conference, it seems clear that RTOP is one of the most often used instruments when studying inquiry instruction and has been seen as the leader in this area over the past several years. A comparative study between RTOP and EQUIP identified many similarities and differences, but the conclusions report that EQUIP has higher inter-rater reliability, provides broader utility (effective and meaningful for both researchers and practitioners), is more targeted to inquiry-based instruction (not the more general constructivist approaches), provides a more comprehensive understanding of teaching performance (macro and micro look at specific teaching aspects), and includes a descriptive rubric to allow teachers and leaders to target tangible goals (Marshall, Smart, Lotter, & Sirbu, 2011).

EQUIP is a valid instrument that was developed, tested, and refined over the past four years and is increasingly being seen as a solid and comprehensive tool for tracking inquiry-based instruction (Marshall, et al., 2010). The framework for EQUIP originated from joining components of many existing frameworks in an attempt to provide a means for studying teachers' transformation toward greater quantity and quality of inquiry-based instruction (Horizon Research, 2002; Llewellyn, 2007; Sampson, 2004; Sawada, et al., 2000).

Although inquiry is not the only student-centered instructional strategy, it is a critical strategy that should be part of every science classroom (Bransford, et al., 2000; NRC, 1996, 2000). For clarity, I use an established definition of inquiry, set forth by *NSES*.

Inquiry is a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations and predictions; and communicating the results. Inquiry requires identification of assumptions, use of critical and logical thinking, and consideration of alternative explanations (NRC, 1996, p. 23).

Various nuances of inquiry are further detailed in the *NSES* and in other research documents and publications (Karplus, 1977; Llewellyn, 2002, 2007; NRC, 2000; NRC, 2012), but the essence of scientific inquiry is clear—students critically and systematically engage in examining, interpreting, and analyzing questions regarding the world around them, and then communicate their findings, providing convincing arguments for their conclusions.

EQUIP OVERVIEW

So how do we assess the quality of the inquiry that teachers lead in their classrooms? How can this assessment be used to improve teacher performance and ultimately student achievement? This chapter provides an overview of the EQUIP instrument, shares the reliability and validity information that EQUIP is based on, offers an example of instruction matched to the scoring, and then shares how

EQUIP provides a predictive indicator of student performance (both content and process). This protocol, which we have named the Electronic Quality of Inquiry Protocol (EQUIP), can be used as a snapshot to measure the quality of inquiry on several indicators for a given class or as a guide that outlines specific areas for teachers to target for growth. EQUIP provides a reliable and valid resource to measure the quality of inquiry that is being facilitated within classrooms. We researched, developed, and refined EQUIP over a period of three years.

Good teachers use many different instructional methods throughout a day, a week, and a year. EQUIP is not designed for all situations; it specifically focuses on the factors associated with the quality of inquiry-based instruction being facilitated by teachers, not with other methods that may be used in the classroom. The complete EQUIP instrument can be accessed from www.clemson.edu/iim. The first three sections of EQUIP are intended mainly for researchers; whereas the remaining sections of EQUIP are useful for both researchers and practitioners alike.

Section 1 contains basic demographic information and descriptive lesson information such as objectives and standards. Section 2 is used to rate five-minute snapshots on several issues such as student attention and cognitive level. Section 3 is used for field notes that teachers may or may not want to use, depending on their goals for individual growth. The column concerning "Classroom Notes of Observation" is for the evaluator to indicate as objectively as possible what has transpired in the class. The "Comments" column is for the evaluator to express her/his interpretation of what is going on; consequently it is more subjective.

Sections 4-8 of EQUIP are critical for researchers and practitioners and should be completed at the culmination of the observed lesson. The quantitative data from these sections of EQUIP include 19 indicators and five different composites. The 19 indicators are divided into the following four constructs:

- **Instruction (5 items):** Section 4.
- **Discourse (5 items):** Section 5.
- **Assessment (5 items):** Section 6.
- **Curriculum (4 items):** Section 7.

After scoring each indicator, five composite scores are generated—Section 8 (one for each construct plus an overall score for the lesson). Each indicator and composite score can range from 1-4 (Level 1 = Pre-inquiry, Level 2 = Developing Inquiry, Level 3 = Proficient Inquiry, and Level 4 = Exemplary Inquiry). The composite scores are based on the essence for that composite rather than an average of all indicators in the composite. I encourage teachers to avoid becoming defensive about the ratings; it is more important to understand why a score falls into a specific level and what can be done to advance to a higher level in the future then to argue over a particular score.

Once a benchmark measurement is determined, teachers can begin, individually or in teams, to chart the growth and target areas where improvement is desired. The descriptive nature of EQUIP helps teachers move from the "I know it when I see it" to an understanding of the specific aspects that were or were not effective. The insights gained when using EQUIP can provide a foundation for developing a plan that will ultimately improve instruction and student learning. See Table 1 for uses of EQUIP.

EQUIP RELIABILITY AND VALIDITY INFORMATION

EQUIP was created in response to a need for a reliable and valid instrument to assess the quantity and quality of inquiry in K-12 math and science classrooms (Marshall, et al., 2010). None of the other protocols met our specific needs for guiding teachers as they plan and implement inquiry-based instruction and for assessing the quantity and quality of inquiry instruction. EQUIP's structure provides both a formative and summative means

Table 1. Various uses of EQUIP paired with intended audience

Use of EQUIP	Teacher	Instructional Leader	Researchers
Establish benchmarks and then chart growth over time	X	X	X
Work with teachers to target growth in performance		X	X
Reflect back upon a lesson	X		
Videotape lesson and then complete protocol either alone or with peers during a replay of the lesson	X	X	X
Complete the instrument while observing another teacher's class	X	X	
Guide conversations with a teacher or team of teachers		X	X

to study inquiry-based instruction in K-12 science and math classrooms. EQUIP was specifically designed to (1) evaluate teachers' classroom practice, (2) evaluate PD program effectiveness, and (3) guide reflective practitioners as they try to increase the quantity and quality of inquiry. Though EQUIP is designed to measure both quantity and quality of inquiry instruction, the reliability and validity issues associated with only the quality of inquiry are addressed below.

Face validity: Five science education researchers, four math education researchers, and two doctoral students in Curriculum and Instruction from four universities helped assess the face validity of EQUIP. Further, two measurement experts with knowledge of instrument development assessed the instrument structure. To guide face validity conversations, we posed the following questions. Does EQUIP seem well-designed way to assess the quality of inquiry? Does it seem as though it will provide reliable measures? For the content specialists, does it maintain fidelity to the discipline (math/science)? Does each indicator, along with descriptor, provide a critical measure that seamlessly progresses from non-inquiry to exemplary inquiry? Finally, does a Level 3 descriptor provide an accurate benchmark representation of proficiency for a given indicator? Through a series of face-to-face meetings, email communication, and phone conversations, each indicator with the accompanying descriptor was scrutinized until both educational researchers and individuals conducting measurements in the field achieved consensus. Negotiation and refinements centered on balancing what theory suggested with what was consistently measurable.

Internal consistency: EQUIP indicators were examined for internal consistency using Cronbach's Alpha (α) for 102 class observations used for field-testing. The α-value ranged from .880-.889, demonstrating strong internal consistency. Further, the indicators that comprise the instrument hold together well in both science and mathematics settings.

Inter-rater reliability: Cohen's Kappa (κ) was used to determine inter-rater reliability from 16 paired observations. Using the Landis and Koch (1977) interpretative scale, κ scores averaged .6 and thus fall between moderate and substantial agreement. Further, since EQUIP items fall along a continuum and not an absolute discrete scale, it is not surprising that Kappa scores were not close to 1.0, which would show total agreement for all items. For these 16 paired observations, the coefficient of determination, r^2, was .856, indicating a strong collective agreement between raters. Specifically, 85.6% of Observer B's assessment is explained by Observer A's assessment and vice versa.

Content and construct validity: Once face validity and high reliability had been established, content validity was examined to provide a deeper analysis of the validity surrounding the instrument.

In assessing content validity, we are essentially asking: How well does EQUIP represent the domain it is designed to represent? In this instance, EQUIP was designed to represent components associated with the quality of inquiry, as defined by the research literature. In order to establish content validity, the primary constructs measures in EQUIP were aligned with key literature associated with inquiry-based instruction. Since only the indicators that remain in the model will be justified with research literature, content validity and construct validity are addressed together.

In evaluating construct validity, a Confirmatory Factor Analysis (CFA) using Structural Equation Modeling (SEM) was ran with model trimming used to eliminate any indicators that did not contribute significantly to each construct. The resulting more parsimonious model, trimmed the 26 total indicators to 19 (five for Instruction, four for Curriculum, five for Discourse, and five for Assessment).

Final EQUIP model: The five indicators (with the theory and research to justify) that comprise the Instructional Factors include: (1) *Instructional Strategies* (Abell & Lederman, 2007; Bransford, et al., 2000; Chiappetta & Koballa, 2006; National Research Council, 2000), (2) *Order of Instruction* (Abell & Lederman, 2007; Biggs, 1996; Bybee, et al., 2006), (3) *Teacher Role* (Lampert, 1990; Mortimer & Scott, 2003; National Research Council, 1996; van Zee, Iwasyk, Kurose, Simpson, & Wild, 2001), (4) *Student Role* (Cobb, Wood, & Yackel, 1990), and (5) *Knowledge Acquisition* (Chinn & Brewer, 1998; Mortimer & Scott, 2003). Note that all four constructs that frame the EQUIP has been thoroughly discussed and validated in prior work (Marshall, 2009). The descriptive rubric used to measure all five Instructional Factor indicators along with the 14 other indicators is found at www.clemson.edu/iim.

The four indicators that were identified in the CFA that comprised the Curriculum Factor construct, along with literature to support its inclusion, include: (1) *Content Depth* (Schmidt, McNight, & Raizen, 2002; Wiggins & McTighe, 1998), (2) *Learner Centrality* (Donovan & Bransford, 2005; Knowles & Brown, 2000; NBPTS, 2000; NRC, 1996), (3) *Integration of Content and Investigation* (Llewellyn, 2002, 2007; Luft, Bell, & Gess-Newsome, 2008; NRC, 2000), and (4) *Organizing and Recording Information* (Marzano, Pickering, & Pollock, 2001).

The five tightly aligned indicators identified as part of the Discourse construct include: (1) *Questioning Level* (Krathwohl, 2002; Vygotsky, 1978),(2) *Complexity of Questions* (Chin, 2007), (3) *Questioning Ecology* (Morge, 2005; Mortimer & Scott, 2003), (4) *Communication Pattern* (Kelly, 2007; Lemke, 1990; Moje, 1995), and (5) *Classroom Interaction* (Lampert, 1990; van Zee, et al., 2001).

The final construct, Assessment, contains the following five indicators: (1) *Prior Knowledge* (Bransford, et al., 2000; Chambers & Andre, 1997), (2) *Conceptual Development* (Driver, Squires, Rushworth, & Wood-Robinson, 1994), (3) *Student Reflection* (Mezirow, 1990; White & Frederiksen, 1998, 2005; Wiggins & McTighe, 1998), (4) *Assessment Type(s)* (Black, Harrison, Lee, Marshall, & Wiliam, 2004; Black & Wiliam, 1998), and (5) *Role of Assessing* (Bell & Cowie, 2001; Stiggins, 2005; Stigler & Hiebert, 1999).

Three indicators were trimmed from the original model even though their importance in quality instruction can be easily justified. It was determined that these three items are too difficult to measure accurately by a single item during a single class period observation and include the following items: (1) *Teacher Content Knowledge*, (2) *Meaningful Context*, and (3) *Fundamental Ideas*.

Although absolute parameters for SEM do not readily exist, the values and justifications for the model include the following: χ^2 is significant $p < .001$, $\chi^2/df \leq 2$ indicates reasonable fit (Kline, 2005), RMSEA of .1 is on the threshold of reasonable fit (Browne & Cudeck, 1993), SRMR < .1 is considered favorable (Kline, 2005), and the computerized fit index, CFI, of > .90 is considered a good fit (Hu & Bentler, 1999). The four-construct model, 19-indicator model, provides a good-fitting model that also is solidly supported by the literature base regarding effective inquiry instruction.

SAMPLE LESSON ANALYZED USING EQUIP

In the following pages, the factors of Instruction, Discourse, Assessment, and Curriculum are discussed in more depth as I illustrate the application of EQUIP. Again, the indicators for these factors are all assessed at the end of the observational period. The example comes from a physical science lesson framed by the essential question *"What factors affect the motion of an object?"* (Portions of the example and discussion that follows were published in *The Science Teacher* (Marshall, Horton, & White, 2009)). In the observation from which the example is drawn, the teacher provided teams of 3 or 4 students with mousetrap racer kits and challenged them to create a mousetrap racer that would go 5 meters the fastest, but would stop before it had traveled 6 meters. This competition incorporated process skills (e.g., asking good scientific questions, collecting meaningful data, analyzing results), and conceptual ideas (e.g., speed, motion, force, conservation of energy) from science, math, and engineering.

Instruction: Table 2 shows two of five indicators that comprise the factor of Instruction. I discuss the ratings for the science observation for the example described above. Because the indicators are associated with the same factor, there are connections among them. However, these connections are not absolute; there are sufficient distinctions among the indicators so that the levels often vary considerably even within the same factor.

Because the teacher provided the vehicle assembly instructions before students had sufficient time to think through their own creation and because she stopped and lectured about the terminology associated with motion, the *Instructional Strategies* earned a Level 2 inquiry rating. Had the teacher provided more opportunities for input of student ideas throughout the investigation, then the quality of the inquiry would have been at least Level 3.

The teacher did, however, achieve a Level 3 inquiry rating for *Order of Instruction* because the lesson engaged the students in exploring concepts before the teacher explained them, and students were involved in explaining their conceptual ideas to the teacher and their peers.

Discourse: Discourse measures the classroom climate and interactions relating to inquiry instruc-

Table 2. Sample of EQUIP instructional indicators associated with inquiry-based instruction

Indicator Measured	*Pre-Inquiry (Level 1)*	*Developing Inquiry (Level 2)*	*Proficient Inquiry (Level 3)*	*Exemplary Inquiry (Level 4)*
Instructional Strategies	Teacher predominantly lectured to cover content.	Teacher frequently lectured and/or used demonstrations to explain content. Activities were verification only.	Teacher occasionally lectured, but students were engaged in activities that helped develop conceptual understanding.	Teacher occasionally lectured, but students were engaged in investigations that promoted strong conceptual understanding.
Order of Instruction	Teacher explained concepts. Students either did not explore concepts or did so only after explanation.	Teacher asked students to explore concept before receiving explanation. Teacher explained.	Teacher asked students to explore before explanation. Teacher and students explained.	Teacher asked students to explore concept before explanation occurred. Though perhaps prompted by the teacher, students provided the explanation.

tion and learning. Two of five indicators associated with this factor are shown in Table 3.

As the lesson progressed, the teacher provided challenging, higher-level questions (e.g., How did your results compare with those from other groups?) as students presented their findings, which resulted in a Level 3 inquiry rating for *Questioning Level*. However, once students responded to the higher-level questions, the quality of the interactions dropped as the teacher followed-up responses with only low-level probes (e.g., How did you find the second point on the graph?). This resulted in a rating of Level 2 for *Classroom Interaction*. The teacher could raise this score by following-up student responses with more thought-provoking questions such as, "Why was the slope calculated by group 2 larger than the slope calculated by group 1? What does that slope tell us?"

Assessment: Five indicators are used to measure the Assessment factor relating to instructional practice. Two of the indicators are shown in Table 4.

Because the teacher did not attempt to assess or take into consideration the prior knowledge students possessed, the lesson earned a Level 1 inquiry rating for *Prior Knowledge*. A short pretest, a KWL chart, or even a discussion concerning what students already knew may have revealed strengths or, on the other hand, some misconceptions regarding motion that should be addressed. The teacher also fell short on *Conceptual Development*. When formative assessments are implemented throughout the lesson, student-learning increases. By making the lesson more prescribed than necessary, critical thinking was minimized. This resulted in a Level 2 rating for this indicator.

Table 3. Sample of EQUIP discourse indicators associated with inquiry-based instruction

Indicator Measured	*Pre-Inquiry (Level 1)*	*Developing Inquiry (Level 2)*	*Proficient Inquiry (Level 3)*	*Exemplary Inquiry (Level 4)*
Questioning Level	Questioning rarely challenged students above the remembering level.	Questioning rarely challenged students above the understanding level.	Questioning challenged students up to application or analysis levels.	Questioning challenged students at various levels, including at the analysis level or higher; level was varied to scaffold learning.
Classroom Interaction	Teacher accepted answers, correcting when necessary, but rarely followed-up with further probing.	Teacher or another student occasionally followed-up student response with further low-level probe.	Teacher or another student often followed-up response with engaging probe that required student to justify reasoning or evidence.	Teacher consistently and effectively facilitated rich classroom dialogue where evidence, assumptions, and reasoning were challenged by teacher or other students.

Table 4. Sample of EQUIP assessment indicators associated with inquiry-based instruction

Indicator Measured	*Pre-Inquiry (Level 1)*	*Developing Inquiry (Level 2)*	*Proficient Inquiry (Level 3)*	*Exemplary Inquiry (Level 4)*
Prior Knowledge	Teacher did not assess student prior knowledge.	Teacher assessed student prior knowledge but did not modify instruction based on this knowledge.	Teacher assessed student prior knowledge and then partially modified instruction based on this knowledge.	Teacher assessed student prior knowledge and then modified instruction based on this knowledge.
Conceptual Development	Teacher encouraged learning by memorization and repetition.	Teacher encouraged product- or answer-focused learning activities that lacked critical thinking.	Teacher encouraged process-focused learning activities that required critical thinking.	Teacher encouraged process-focused learning activities that involved critical thinking that connected learning with other concepts.

When students are challenged to defend their solutions to scientific questions, a Level 3 or 4 rating is appropriate.

Curriculum: The EQUIP includes four indicators associated with various Curriculum issues related to inquiry instruction. These indicators are tied directly to what is experienced by students, not what appears in a text or notes. *Organizing and Recording Information* is one of several areas in which teachers can provide students with different levels of scaffolding—thus differentiating instruction. The goal is to challenge all students to their highest level while not overly frustrating anyone. For instance, one student with a learning disability may need the structure that a graphic organizer provides, whereas an ESL student may need more visuals to help decode the language barriers. We should always strive to help students progress to a level where less direct assistance is needed. By doing so, we will have encouraged and helped to develop habits of lifelong learning. To earn Level 4 on this and other indicators, teachers should consider the various needs of *all* students in their class. Two of the Curriculum indicators, *Integration of Content and Investigation* and *Organizing and Recording Information*, are displayed in Table 5.

The *Integration of Content and Investigation* earned a Level 3 inquiry rating because the investigation almost continually integrated concepts such as speed vs. time graphs and conservation of energy into the student investigations. *Organizing and Recording Information* was scored at Level 2 because the teacher provided little opportunity for the students to determine how the data should be collected and organized. When data sheets are provided with the headings and axes already labeled, which is what happened during this observation, students are deprived of a rich opportunity to think about how to collect, organize, and convey meaning from the data. By having the opportunity to organize and record information as they see fit, students think more deeply and more critically about the concepts being investigated (e.g., how many trials are needed? Is speed the independent or dependent variable, and why?). Had the teacher provided this opportunity, the rating for this indicator would have risen to a Level 3 or 4.

IMPROVING QUALITY OF INQUIRY TEACHING

After each of the indicators associated with the four factors has been assessed, in Section 8 of EQUIP an overall, or holistic, rating is determined for each factor. Again, this holistic rating is not necessarily the mean of the indicators, but is the Level that best captures the essence of the lesson. Though it may seem that approaching the rating this way would make this section overly subjective, we have found that our inter-rater reliability, or consistency between different raters, is quite high.

Table 5. Sample of EQUIP curriculum indicators associated with inquiry-based instruction

Indicator Measured	*Pre-Inquiry (Level 1)*	*Developing Inquiry (Level 2)*	*Proficient Inquiry (Level 3)*	*Exemplary Inquiry (Level 4)*
Integration of Content & Investigation	Lesson either content-focused or activity-focused but not both.	Lesson provided poor integration of content with activity or investigation.	Lesson incorporated student investigation that linked well with content.	Lesson seamlessly integrated the content and the student investigation.
Organizing & Recording Information	Students organized and recorded information in prescriptive ways.	Students had only minor input as to how to organize and record information.	Students regularly organized and recorded information in non-prescriptive ways.	Students organized and recorded information in non-prescriptive ways that allowed them to effectively communicate their learning.

Once the instrument has been completed and the current state of inquiry instruction is established, the next step is to improve the quality of inquiry. Though establishing the benchmark may bring about some change just by having specific aspects of instructional practice brought to the teacher's attention, the goal is to become more intentional and explicit by developing an action plan of next steps. It is normal to desire to improve everything that ails our instruction all at once. However, such a course of action often leads to frustration and undue anxiety; effective change is usually incremental.

I recommend that teachers focus on one specific indicator that they wish to improve upon during the next lesson or unit of study. Once the desired growth has been achieved, then it is time to tackle another indicator. After four indicators relating to inquiry instruction have been improved, perhaps one from each of the factors Instruction, Discourse, Assessment, and Curriculum, the teacher should strive to maintain that level of performance before undertaking more improvements. If teachers work together and note common areas for growth, it may make sense to work on certain indicators together. This shared approach provides a support structure to exchange thoughts and ideas.

If current practice falls largely in Level 1, then it makes sense to begin reading about constructivist approaches to learning and inquiry-based methods of teaching, looking for examples of lessons and instruction. Many articles from NCTM journals (e.g. *The Science Teacher*, *Science Scope*), along with journals from other professional organizations, provide these, along with many innovative ideas that can be of immense value to teachers who wish to modify their practice. In addition, teachers may seek out any one of many professional development institutes that provide opportunities to experience inquiry learning firsthand. Generally, a Level 2 performance suggests that a teacher is familiar with getting students engaged and active, but the lessons tend to be more prescriptive, with students having only limited opportunities to develop the ideas for themselves. Additionally, instruction is still heavily teacher-focused. At Level 3, the teacher has demonstrated a student-centered inquiry-learning environment that actively engages students in investigations, questioning, and explanations. The role of the teacher remains vital (as it does at all levels), but she now functions more as a facilitator who scaffolds learning experiences than as a giver of facts and knowledge.

I do not expect that any one lesson would merit a Level 4 for all indicators or even for all factors. In fact, I have yet to see such a lesson, and I have seen some amazing lessons. Further, the point is not to make every instructional moment a Level 3 or higher; rather, the goal is to help teachers become more intentional about their practice. By making teachers aware of what high quality inquiry practice entails, I believe they will be more likely to implement it successfully when it is their desired instructional approach.

EQUIP provides teachers with a concrete way to reflect on their own teaching practice as they strive to lead inquiry-based learning experiences in their classroom. Inquiry instruction is challenging to implement well, but, when done effectively, learning is clearly evident with *all* students and at *all* ability levels.

EQUIP AND STUDENT ACHIEVEMENT

The value of EQUIP is partially predicated on the fact that it provides a reliable, valid metric to measure of inquiry-based instruction. In and of itself, this is valuable, but the ultimate goal is to find an instrument that helps predict student achievement based on teacher performance. Thus, the challenge becomes finding the most appropriate dependent measure for student achievement. Specifically, what metric would provide valuable student achievement growth data with minimal testing bias?

The Measures of Academic Progress (MAP) from Northwest Evaluation Association (NWEA) was selected as the metric to study student growth in science classrooms. MAP, a reliable and valid assessment (NWEA, 2004) used by schools in 48 states, is an adaptive test that provides either more or less challenging items, depending on students' success or failure on previous questions. Further, because it is aligned with state and national science standards (NWEA, 2005), MAP can pinpoint students' current level of achievement. Students are assessed both in the fall and in the spring; hence, growth during the majority of the academic year can be readily determined. MAP also allows success to be studied with various ability levels, thus providing a means to research possible effects on the achievement gap.

The MAP test provides a mechanism to measure the growth of students in both science content and science process during the course of a given academic year for elementary through early high school grades. Further, in a collaborative effort with NWEA (funding by an NSF grant), I was able to compare the growth of the students in my study group to similar students from around the state who also took the test, the Virtual Comparison Group (VCG). The VCG is comprised of students who: 1) are the same race and gender, 2) took the test during the same test window (both pre- and post-), and 3) attend a school with a similar free and reduced lunch status (a measure of poverty in the school). Each of the students in the intervention is compared to 21-51 like students from the VCG.

MAP inherent strengths are several. First, because test items are aligned to state science standards, it has high predictive validity when compared to other state assessments (Cronin, Kingsbury, Dahlin, Adkins, & Bowe, 2007; NWEA, 2005). Second, performance of a given teacher's students can be studied at both in terms of science content and process knowledge scores without requiring an additional test. Third, because it is adaptive, MAP provides a broader, more robust sample of the entire domain than a fixed-form test does (NWEA, 2003). Finally, since the districts that I work with already use MAP, no additional testing is necessary to obtain a reliable and valid measure of student performance.

With EQUIP as the predictor measure (teacher performance) and MAP scores as the dependent measure (student achievement), a correlation was analyzed to see if teacher performance relative to inquiry-based instruction could partially predict student achievement growth during the academic year. The results seemed mixed at first, but then a clear pattern emerged. Specifically, it was found that EQUIP could not predict for student growth in content knowledge or process skills during the first year of involvement in transformation of practice—no significance found. This lack of significance was attributed to several reasons: 1) teacher performance is unstable during times of transformation (larger variation of practice), 2) teachers tend to "perform" for observers (evaluators) during the first year often providing what they think is expected but which is often atypical of most of their instruction, and 3) inquiry teaching is an inconsistent part of everyday practice.

During the second year of involvement with teachers through a long-term professional development training program, it was found that teachers' changes in practice became more solidified and consistent. In addition to instruction becoming more consistent relative to inquiry, EQUIP became a powerful predictor of the growth seen in students during the academic year. Specifically, the average overall lesson score earned on the EQUIP for a teacher over the course of four observations, was able to explain approximately 36% of variance in the student growth. These data will continue to be analyzed as more teachers become involved in the second year portion of the program. However, the initial two years of data indicate a very power metric that could be used to improve student achievement in both science content and process. It would be expected that inquiry-based instruction increases students' performance on science process knowledge, but these initial results suggest that the effect is equally high for science content knowledge as well.

IMPLICATIONS

In an effort to help teachers move beyond an "I know it when I see it" mentality regarding inquiry, I propose the use of EQUIP as an instrument to help guide teacher practice to greater quantity and quality of inquiry-based instruction. The evidence suggests that when high-quality, proficient inquiry (Level 3 on EQUIP) becomes a consistent portion of instruction then the growth in student achievement exceeds a Virtual Comparison Group (VCG) of typical instructional practice. It is infrequent in educational research to find strong predictive indicators of things that when done by the teacher in the classroom result in improved student achievement.

Further, it is exciting to confirm that when inquiry is implemented well in the classroom that student achievement exceeds the comparison group on process skill knowledge (e.g., interpreting a graph or designing a study). However, these findings also show that we do not need to just "teach to the test" (e.g., tell, rehearse, memorize), because as students improve process knowledge, it is also possible for them to improve their content knowledge at a rate that also significantly exceeds the VCG.

Obviously, EQUIP is not a single panacea for all that ails our education system, but it does provide a clear, descriptive means to guide the transformation of instructional practice so that student learning is greatly increased via inquiry-based forms of instruction. Because of the complex, multifaceted nature of inquiry instruction, it has been very challenging to develop a protocol that assesses the quality of inquiry instruction in a valid and reliable manner. EQUIP seems to have met that challenge and was designed to (1) evaluate teachers' classroom practice, (2) evaluate PD program effectiveness, and (3) provide a tool to guide reflective practitioners as they strive to increase the quantity and quality of inquiry that they lead in their classrooms (Marshall, Horton, & White, 2009). The culminating four-construct (Instruction, Curriculum, Discourse, and Assessment) EQUIP is a reliable and valid instrument that meets these goals.

ACKNOWLEDGMENT

This material is based upon the work supported by the National Science Foundation under Grant #DRL-0952160 and from a grant from the South Carolina Commission on Higher Education under the auspices of the EIA Teacher Education Centers of Excellence Grant Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Abell, S. K., & Lederman, N. G. (2007). *Handbook of research on science education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Adams, P. E., & Krockover, G. H. (1999). Stimulating constructivist teaching styles through use of an observation rubric. *Journal of Research in Science Teaching*, *36*(8), 955–971. doi:10.1002/(SICI)1098-2736(199910)36:8<955::AID-TEA4>3.0.CO;2-3

Beerer, K., & Bodzin, A. (2003). *Science teacher inquiry rubric (STIR)*. Retrieved April 25, 2007, from http://www.lehigh.edu/~amb4/stir/stir.pdf

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*, 536–553. doi:10.1002/sce.1022

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. doi:10.1007/BF00138871

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 9–21.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74. doi:10.1080/0969595980050102

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A., & Long, J. S. (Eds.), *Testing Structural Equation Models* (pp. 136–162). Beverly Hills, CA: Sage.

Burry-Stock, J., & Oxford, R. (1994). Expert science teaching edcuational evaluation model (ESTEEM): Measuring excellence in science teaching for professional development. *Journal of Personnel Evaluation in Education*, *8*(3), 267–297. doi:10.1007/BF00973725

Bybee, R. W., Taylor, J. A., Gardner, A., Scotter, P. V., Powell, J. C., & Westbrook, A. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications*. Colorado Springs, CO: BSCS.

Capps, D. K., Crawford, B. A., & Constas, M. A. (2012). A review of empirical literature on inquiry professional development: Alignment with best practices and a critique of the findings. *Journal of Science Teacher Education*, *23*(3), 291–318. doi:10.1007/s10972-012-9275-2

Chambers, S. K., & Andre, T. (1997). Gender, prior knowledge, interest and experience in electricity and conceptual change text manipulations in learning about direct current. *Journal of Research in Science Teaching*, *34*(2), 107–123. doi:10.1002/(SICI)1098-2736(199702)34:2<107::AID-TEA2>3.0.CO;2-X

Chiappetta, E. L., & Koballa, T. R. J. (2006). *Science instruction in the middle and secondary schools: Developing fundamental knowledge and skills for teaching* (6th ed.). Upper Saddle River, NJ: Pearson Perrill Prentice Hall.

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching*, *44*(6), 815–843. doi:10.1002/tea.20171

Chinn, C. A., & Brewer, W. F. (1998). Theories of knowledge acquisition. In Fraser, B. J., & Tobin, K. (Eds.), *International Handbook of Science Education* (pp. 97–113). London, UK: Kluwer Academic Publishers.

Cobb, P., Wood, T., & Yackel, E. (1990). Classrooms as learning environments for teachers and researchers. In Davis, R. B., Maher, C. A., & Noddings, N. (Eds.), *Constructivist Views of the Teaching and Learning of Mathematics* (pp. 125–146). Reston, VA: NCTM. doi:10.2307/749917

Cronin, J., Kingsbury, G. G., Dahlin, M., Adkins, D., & Bowe, B. (2007). *Alternate methodologies for estimating state standards on a widely-used computerized adaptive test*. Paper presented at the National Council on Measurement in Education. New York, NY.

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan*, *92*(6), 81–92.

Donovan, M. S., & Bransford, J. D. (2005). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: National Academies Press.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas*. London, UK: Taylor & Francis Ltd.

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. *American Educational Research Journal*, *38*(4), 915–945. doi:10.3102/00028312038004915

Horizon Research. (2002). *Inside the classroom interview protocol*. Retrieved from http://www.horizon-research.com/instruments/clas/cop.php

Hu, L., & Bentler, P. M. (1999). Cutoff criteria in fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Karplus, R. (1977). Science teaching and the development of reasoning. *Journal of Research in Science Teaching*, *14*, 169–175. doi:10.1002/tea.3660140212

Kelly, G. J. (2007). Discourse in science classrooms. In Abell, S. K., & Lederman, N. G. (Eds.), *Handbook of Research on Science Education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Knowles, T., & Brown, D. F. (2000). *What every middle school teacher should know*. Portsmouth, NH: Heinemann.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. doi:10.1207/s15430421tip4104_2

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal*, *27*(1), 29–63.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi:10.2307/2529310

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex.

Llewellyn, D. (2002). *Inquiry within: Implementing inquiry-based science standards*. Thousand Oaks, CA: Corwin Press.

Llewellyn, D. (2007). *Inquiry within: Implementing inquiry-based science standards in grades 3-8* (2nd ed.). Thousand Oaks, CA: Corwin Press.

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

Luft, J., Bell, R. L., & Gess-Newsome, J. (2008). *Science as inquiry in the secondary setting*. Arlington, VA: National Science Teachers Association.

Marshall, J. C. (2009). *The creation, validation, and reliability associated with the EQUIP (electronic quality of inquiry protocol): A measure of inquiry-based instruction*. Paper presented at the National Association of Researchers of Science Teaching Conference. Washington, DC.

Marshall, J. C., Horton, B., & Edmondson, E. (2007). *4E x 2 instructional model*. Retrieved from http://www.clemson.edu/iim

Marshall, J. C., Horton, B., Igo, B. L., & Switzer, D. M. (2009). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom. *International Journal of Science and Mathematics Education*, *7*(3), 575–596. doi:10.1007/s10763-007-9122-7

Marshall, J. C., Horton, B., & White, C. (2009). EQUIPping teachers: A protocol to guide and improve inquiry-based instruction. *Science Teacher (Normal, Ill.)*, *76*(4), 46–53.

Marshall, J. C., & Horton, R. M. (2009). *Developing, assessing, and sustaining inquiry-based instruction: A guide for math and science teachers and leaders*. Saarbruecken, Germany: VDM Publishing House Ltd.

Marshall, J. C., Smart, J., & Horton, R. M. (2010). The design and validation of EQUIP: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education*, *8*(2), 299–321. doi:10.1007/s10763-009-9174-y

Marshall, J. C., Smart, J., Lotter, C., & Sirbu, C. (2011). Comparative analysis of two inquiry observational protocols: Striving to better understand the quality of teacher facilitated inquiry-based instruction. *School Science and Mathematics*, *111*(6), 306–315. doi:10.1111/j.1949-8594.2011.00091.x

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: ASCD.

Mezirow, J. (1990). *Fostering critical reflection in adulthood: A guide to transformative and emancipatory learning*. San Francisco, CA: Jossey-Bass.

Moje, E. B. (1995). Talking about science: An interpretation of the effects of teacher talk in a high school classroom. *Journal of Research in Science Teaching*, *32*(4), 349–371. doi:10.1002/tea.3660320405

Morge, L. (2005). Teacher-pupil interaction: A study of hidden beliefs in conclusion phases. *International Journal of Science Education*, *27*(8), 935–956. doi:10.1080/09500690500068600

Mortimer, E. F., & Scott, P. H. (2003). *Meaning making in secondary science classrooms*. Maidenhead, UK: Open University Press.

National Academy of Sciences. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press.

National Board for Professional Teaching Standards. (2000). *A distinction that matters: Why national teacher certification makes a difference*. Greensboro, NC: Center for Educational Research and Evaluation.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

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

National Science Teachers Association. (1998). *Standards for science teacher preparation*. Retrieved from http://www.nsta.org/main/pdfs/nsta98standards.pdf

Northwest Evaluation Association. (2003). *Technical manual*. Lake Oswego, OR: Northwest Evaluation Association.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level tests and measure of academic progress*. Retrieved from http://www.nwea.org

Northwest Evaluation Association. (2005). *NWEA reliability and validity estimates: Achievement level tests and measures of academic progress*. Lake Oswego, OR: Northwest Evaluation Association.

Riggs, I. M., & Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education*, *74*(6), 625–637. doi:10.1002/sce.3730740605

Sampson, V. (2004). The science management observation protocol. *Science Teacher (Normal, Ill.)*, *71*(10), 30–33.

Sawada, D., Piburn, M., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). *Reformed teaching observation protocol (RTOP).* Technical Report No. IN00-01. Phoenix, AZ: Arizona State University

Schmidt, W. H., McNight, C. C., & Raizen, S. A. (2002). *A splintered vision: An investigation of U.S. science and mathematics education*. Retrieved from http://imc.lisd.k12.mi.us/MSC1/Timms.html

Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, *87*(4), 324–328.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: The Free Press.

van Zee, E. H., Iwasyk, M., Kurose, A., Simpson, D., & Wild, J. (2001). Student and teacher questioning during conversations about science. *Journal of Research in Science Teaching*, *38*(2), 159–190. doi:10.1002/1098-2736(200102)38:2<159::AID-TEA1002>3.0.CO;2-J

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*(1), 3–118. doi:10.1207/s1532690xci1601_2

White, B. Y., & Frederiksen, J. R. (2005). A theoretical framework and approach for fostering metacognitive development. *Educational Psychologist*, *40*(4), 211–223. doi:10.1207/s15326985ep4004_3

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: ASCD.