JEFF C. MARSHALL, JULIE SMART and ROBERT M. HORTON

THE DESIGN AND VALIDATION OF EQUIP: AN INSTRUMENT TO ASSESS INQUIRY-BASED INSTRUCTION

Received: 24 February 2009; Accepted: 25 June 2009

ABSTRACT. To monitor and evaluate program success and to provide teachers with a tool that could support their transformation in teaching practice, we needed an effective and valid protocol to measure the quantity and quality of inquiry-based instruction being led. Existing protocols, though helpful, were either too generic or too program specific. Consequently, we developed the Electronic Quality of Inquiry Protocol (EQUIP). This manuscript examines the 2-year development cycle for the creation and validation of EQUIP. The protocol evolved over several iterations and was supported by validity checks and confirmatory factor analysis. The protocol's strength is further supported by high internal consistency and solid interrater agreement. The resulting protocol assesses 19 indicators aligned with four constructs: instruction, curriculum, assessment, and discourse. For teachers, EQUIP provides a framework to make their instructional practice more intentional as they strive to increase the quantity and quality of inquiry instruction. For researchers, EQUIP provides an instrument to analyze the quantity and quality of inquiry being implemented, which can be beneficial in evaluating professional development projects.

KEY WORDS: EQUIP, inquiry, inquiry-based instruction, inquiry protocol, mathematics education, observational protocol, professional development, professional development protocol, science education

INTRODUCTION

The call to align science and mathematics instruction with reform-based initiatives that focus intensely on inquiry-based instructional practices has been met with varying degrees of success. The *National Science Education Standards* (NSES; National Research Council, NRC, 1996) and the *Principles and Standards for School Mathematics* (PSSM; National Council of Teachers of Mathematics, NCTM, 2000) along with many other reform documents (American Association for the Advancement of Science, 1993, 1998; Bransford, Brown, & Cocking, 2000; Llewellyn, 2002; NCTM, 1991; NRC, 2000) state that inquiry-based instruction should be a central tenet of sound instructional practice. However, merely increasing the quantity of inquiry instruction is not sufficient; the quality of inquiry instructional practice must be at such a level that teachers are effective in facilitating rigorous, standard-based, and inquiry-based learning.

Currently, there is little consistency in how science and math teachers describe, understand, and implement high-quality inquiry-based instruction (Marshall, Horton, Igo, & Switzer, 2009; Marshall, Horton, & Smart, 2009). Without guidance indicating otherwise, many educators believe that simply engaging students in activities defines successful inquiry instruction (Moscovici & Holdlund-Nelson, 1998). Other educators see successful inquiry as a deep investigation of the process skills even when no essential content is being explored. Thus, conceptions are often disconnected from the vision communicated by reform-based documents such as NSES. Until clear direction is provided for educators at all levels, the call for transformation to inquiry-based practice will garner mixed results at best.

This article details the development and validation of the Electronic Quality of Inquiry Protocol (EQUIP), 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. Though other protocols provide valuable assistance to educators, none 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. Our research sought to provide one viable mechanism, or protocol, that can be used to assess critical constructs associated with inquiry-based instruction. Our expectation is that this protocol will provide both a formative and summative means to study inquiry-based instruction in K-12 science and math classrooms. Further, we hope that the protocol can be used to guide pre- and in-service teachers' discussions and analyses of inquiry-based instruction.

Review of Literature

Inquiry Instruction

In order to measure the quantity and quality of inquiry facilitated in the classroom, we began with an established definition of inquiry, set forth by NSES, to guide our efforts during the development of the instrument.

> 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 (p. 175) and in other research documents and publications (Karplus, 1977; Llewellyn, 2002, 2007; National Research Council, 2000), 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.

We sought an instrument that would help us understand when and to what degree teachers are effectively facilitating inquiry-based learning experiences. Though some other classroom observational protocols emphasize constructivist-based learning, they generally focus more on overall instructional quality. Our needs called for a research-tested valid instrument that focused directly on measuring the constructs associated with inquiry-based instructional practices. Although we sought a model for both science and math education, science provided a stronger research base for inquiry-based models and protocols. Consequently, our development process drew more upon the science literature than the math literature. However, we also knew that a compromise was needed in order to develop an instrument that would be beneficial to teachers and researchers in both math and science.

Rationale and Need for EQUIP Protocol

In our search for a protocol, we found several instruments that all have significant value. However, none of them fully matched our needs.

Inside the Classroom Observational Protocol (Horizon Research, 2002) provides a solid global view of classroom practice. However, in providing such a broad view of instruction, it does not offer the rigorous and granular understanding of inquiry instructional practice that we were seeking.

The *Reformed Teaching Observation Protocol* (RTOP; Sawada, Piburn, Falconer, Turley, Benford & Bloom, 2000) focuses on constructivist classroom issues, but goes beyond a look at inquiry-based instruction to more of an evaluation of teaching. Furthermore, the use of a Likert scale to assess classroom instruction was a limiting factor for our needs. Specifically, though a Likert scale may be helpful to a researcher in quantifying an observation, it is difficult for teachers to know what they need to do to improve from, say, a 4 to a 5. To fill this gap, we sought an instrument with a descriptive rubric that can be used to guide teachers and help them set specific incremental targets as they seek to improve their inquiry-based instruction.

Finally, evidence is lacking to justify using the RTOP in a granular way to examine individual components of practice. This granular view of practice is needed to develop recommendations for teachers as they improve their individual practice. The RTOP can be substantiated at the macro level (e.g., looking at the total score earned), which again may be helpful to a researcher, but an exploratory factor analysis showed that some but not all of the individual items within a given construct loaded together (Piburn & Sawada, 2001; Sawada et al., 2000). This raises some concerns regarding the validity of the instrument.

The *Science Teacher Inquiry Rubric* (Beerer & Bodzin, 2003) provides a brief protocol that is nicely aligned with the NSES definition. However, it was designed to determine whether stated standards were achieved during instruction; it does not provide insight into the specifics of inquiry that teachers must facilitate with each aspect of inquiry. Even though the rubric seems to be the most literally aligned with the NSES definition of inquiry, there was no reliability or validity information addressed in the studies that used it. Further, this literal translation from the definition to the rubric missed an aspect that is critical for us—looking specifically at the teacher practices that encourage inquiry-based learning.

The *Science Management Observation Protocol* (SMOP; Sampson, 2004) emphasizes classroom management issues and the use of time that support effective science instruction. Though appropriate classroom and time management is essential for effective inquiry-based instruction, the SMOP does not assess key components of inquiry-based instruction.

Finally, teacher efficacy scales (Riggs & Enochs, 1990) have been used as a measure to predict whether reform is likely to occur. This approach is often used because self-reports of efficacy have been closely tied to outcome expectancy (Saam, Boone, & Chase, 2000). However, instead of focusing on teacher self-reported efficacy, our need was for an instrument focused on explicit observable characteristics of inquiry that could be reliably measured.

Since our intent was to measure the quantity and quality of inquiry-based instruction that was occurring in the classroom from a very granular view, our needs were only partially addressed by any one of these instruments. Informed by the existing frameworks (Horizon Research, 2002; Llewellyn, 2007; Sampson, 2004; Sawada et al., 2000), we developed the EQUIP. Because we wanted a single valid instrument, we decided to create this new protocol with a unified framework, instead of cropping from multiple instruments (Henry, Murray, & Phillips, 2007).

The aforementioned protocols have provided leadership in the area of instructional observation (Banilower, 2005; Piburn & Sawada, 2001). However, these protocols did not meet our professional development (PD) objectives. Consequently, we created EQUIP so we could assess constructs relevant to the quantity and quality of inquiry instruction facilitated in science and mathematics classrooms. Specifically, EQUIP was 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 in this manuscript.

Instrument Development

Context of Development

As part of a PD program between a major research university and a large high needs school district (over 68,000 students), we desired to see to what degree science and math teachers were successful in implementing rigorous inquiry-based instruction. The goal of the PD program was to transform teacher practice toward greater quantity and quality of inquiry-based instruction. While many instructional models could be used as a framework for planning inquiry-based instruction, the program specifically endorsed the 4E×2 Instructional Model (Marshall et al., 2009). According to the model, student achievement increases when teachers effectively incorporate three critical learning constructs into their teaching practice: (1) inquiry instruction (NRC, 2000), (2) formative assessment (Black & Wiliam, 1998), and (3) teacher reflection (National Board for Professional Teaching Standards, NBPTS, 2006). The 4E×2 Instructional Model integrates these learning constructs into a single dynamic model that is used to guide transformation of instructional practice.

The 4E×2 Instructional Model builds upon the 5E Instructional Model (Bybee, Taylor, Gardner, Scotter, Powell, Westbrook et al., 2006) and other inquiry models (Atkin & Karplus, 1962; Bybee et al., 2006; Eisenkraft, 2003; Karplus, 1977) by integrating inquiry instruction, formative assessment, and teacher reflection into a single cohesive model. To guide and assess teachers' transformation to inquiry-based instruction, we undertook the challenge of developing and validating EQUIP, outlined in Figure 1.

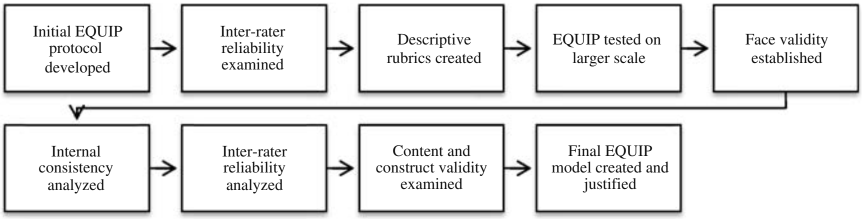
Initial EQUIP protocol developed
Inter-rater reliability examined
Descriptive rubrics created
EQUIP tested on larger scale
Face validity established
Internal consistency analyzed
Inter-rater reliability analyzed
Content and construct validity examined
Final EQUIP model created and justified

Figure 1. Flowchart of the design and validation of EQUIP

Even though we used the EQUIP to measure the effectiveness of teachers in implementing the 4E×2 Instructional Model, EQUIP was designed broadly enough to measure inquiry instruction that does not align with our model.

Development: Semester One

Initial EQUIP Protocol. The development of EQUIP began with two primary steps: (1) determining constructs relevant to the quality of inquiry from the literature and (2) examining existing protocols that aligned with our program goals and with NSES (NRC, 1996) and PSSM (NCTM, 2000) in order to build on previous work in the field. Based on the literature, the following constructs guided early formation of the instrument: instructional factors, ecology/climate, questioning/assessment, and fundamental components of inquiry. The components of inquiry included student exploration before explanation, use of evidence to justify conclusions, and extending learning to new contexts. Even though the RTOP and Inside the Classroom Observational Protocol did not address all of our needs, they did provide some helpful guidance into the types of items for our initial version of EQUIP.

Interrater Reliability. We piloted the initial version of EQUIP in high school science and math classrooms for one academic semester. Our research team (a science education professor, a math education professor, and a curriculum and instruction doctoral student) conducted individual and paired observations in order to assess interrater reliability and validity issues and to clarify operational definitions of constructs. These initial conversations led to preliminary item refinements and pointed toward the need for a more reliable scale of measurement.

Descriptive Rubrics. During these discussions, we realized that a Likert scale did not give us the specific look at the components we wanted and was difficult to interpret until a final summative observational score was rendered. Even then, generalizations about teachers' practice were often difficult to make. Further, the combination of a Likert-scale measure for each item and the summative observational score did not provide the resource we wanted to guide teacher reflection and thus transformation of practice. Specifically, teachers had a difficult time understanding the criteria for each Likert rating and subsequently did not have the formative feedback needed to adjust their practice to align with quality standards of inquiry. Our research team concluded that a descriptive rubric would provide operational definitions of each component of inquiry at various developmental levels.

A descriptive rubric provided several advantages. First, it provided a quantifiable instrument with operationalized indicators. Operationalizing each indicator within the constructs would give EQUIP a more detailed representation of the characteristics of inquiry, allow for assessment of program effectiveness, and provide detailed benchmarks for reflective practitioners. Second, by developing a descriptive rubric, raters would become more systematic and less subjective during observations, thereby bolstering instrument reliability. Finally, the descriptive rubric created that would describe and distinguish various levels of inquiry-based instructional proficiency.

Development: Semesters Two and Three

During the next stage, we worked on creating the descriptive rubrics format for each item that we were assessing with EQUIP. We established four levels of inquiry instruction: pre-inquiry (level 1), developing (level 2), proficient (level 3), and exemplary (level 4). We wrote level 3 to align with the targeted goals laid forth by the science and math standards. Four science education faculty, three math education faculty, and two doctoral students confirmed that all level 3 descriptors measured proficient inquiry-based instructional practice. Llewellyn's work (2005, 2007) also provided an example of how we could operationalize indicators so that they would be of value to both researchers and practitioners.

In addition to the changes in the assessment scale, we reorganized EQUIP to better align the indicators to the major components of instructional practice that could be explicitly observed. The initial protocol targeted three such components: instruction, curriculum, and ecology. Our initial instrument began with these three components because these central tenets repeatedly surface in the literature as the major components of effective instructional practice for both pre- and in-service teachers (Interstate New Teacher Assessment and Support

Consortium (INTASC), 1992). Thus, the protocol would involve the specific aspects within these components that relate to proficient inquiry-based instruction. However, assessing student learning is also a critical piece of the literature; therefore, we decided that it was important to integrate assessment into each of the three components. Later, our statistical analysis of the individual indicators provided clear justification for separating it into its own component.

During the initial stage, our team reviewed items and field-tested the rubrics to see if each level for each item was discrete and observable. We received further input during two state and three national research conferences during follow-up discussions. The combined feedback from these individuals led to further refinement of the descriptive rubric and rewording of items to clarify constructs measured by EQUIP.

Development: Semester Four

After three semesters of development, EQUIP's new seven-section format was ready for more rigorous testing. Sections I–III addressed demographic details (e.g., highest degree earned, number of years teaching, ethnicity, gender breakdown of students), use of time (e.g., activity code, cognitive code, inquiry instruction component), and qualitative notes to provide support and justification of claims made. These sections, however, were not involved in the reliability and validity claims being tested and thus are not addressed in this manuscript.

Sections IV–VI, to be completed immediately after an observation, addressed instruction, curriculum, and ecology. These three constructs assessed 26 total indicators: nine for instruction (e.g., conceptual development, order of instruction), eight for curriculum (e.g., content depth, assessment type), and nine for ecology (e.g., classroom discourse, visual environment). Finally, Section VII provided a summative assessment of time usage, instruction, curriculum, and ecology, and a holistic overall assessment of the inquiry presented in the lesson.

EQUIP Tested on Larger Scale. This version of EQUIP was piloted in middle school science and math classrooms for 5 months. Four raters conducted both paired and individual observations. Raters met immediately after paired observations, and the entire team met weekly to discuss the protocol, our ratings, and challenges we faced. Details regarding the validation of EQUIP are discussed in the next sections.

Instrument Validation

Research Team and Observations

With the addition of another curriculum and instruction doctoral student, our research team grew to four members. The three original members were involved in the initial development and refinement of EQUIP and were therefore familiar with the instrument and its scoring. Our fourth member joined the team at the beginning of the validation period.

Prior to conducting official classroom observations, all team members took part in a video training session where we viewed prerecorded math and science lessons and rated them using EQUIP. Follow-up conversations helped us clarify terminology and points of divergence. Observations from this training were not included in the analyses of reliability and validity.

Our research team then conducted a total of 102 observations, including 16 paired observations, over the next 5 months. Observations occurred in the classrooms of 22 middle school (grades 6–8) math ($n=10$) and science ($n=12$) teachers at two high diversity schools. Free and reduced lunch percentages for the two schools were 38.3% and 56.0% with the respective non-Caucasian populations being 46.7% and 68.0%. All data were entered into Microsoft Access, converted into an Excel spreadsheet, and then used SPSS and Mplus for analysis. A broad range of instructional performance was seen, thus allowing the maximum range of scores (1–4) to be coded for each indicator.

Validity

Face Validity. In addition to the four members on the project, four science education researchers and three math education researchers from three additional universities helped assess the face validity. 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 like a reasonable 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 noninquiry 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 all 102 class observations. The α-value ranged from 0.880–0.889, demonstrating strong internal consistency. For the science observations ($n = 60$), the standardized α-value ranged from 0.869–0.874, and for the math observations ($n = 42$), the range was 0.823–0.861. Thus, the instrument items hold together well as a whole and for science and mathematics separately.

Interrater Reliability. We conducted 16 paired observations to analyze interrater reliability, via Cohen's kappa (κ). The κ scores averaged 0.61 for the nine indicators for instruction, 0.62 for the eight indicators for curriculum, and 0.55 for the nine indicators for ecology. Using the Landis and Koch (1977) interpretative scale, these data fall between moderate and substantial agreement.

For these 16 paired observations, the coefficient of determination, r^2, was 0.856 (see Figure 2). The r^2 value indicates a more collective view of agreement between the raters. Specifically, 85.6% of observer B's assessment is explained by observer A's assessment and vice versa. This value was generated using a summative score that included all 26 indicators plus the five overall ratings for each paired observation. When the observations were separated by middle school science ($n = 9$) and middle school math ($n = 7$), the respective r^2 values were 0.958 and 0.820.

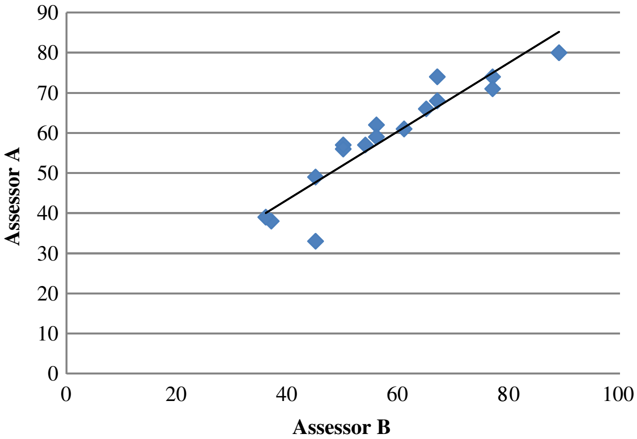

Figure 2. Coefficient of determination between Assessor A and B

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 NSES standards for inquiry and key literature associated with inquiry-based instruction. Since only the factors that remain in the model will be justified with research literature, we address the content validity and construct validity together.

In evaluating construct validity, we ran a confirmatory factor analysis (CFA) on our three constructs (instruction, curriculum, and ecology). CFA was achieved using structural equation modeling (SEM) for the three constructs with model trimming used to eliminate any indicators that did contribute significantly to each construct. In an attempt to achieve the most parsimonious model, the first SEM trimmed the 26 total indicators to 14 (five for instruction, four for curriculum, and five for ecology).

Final EQUIP Model. After confirming internal consistency (α-values ranged from 0.858–0.912), we discussed the content validity of the new three-construct 14-indicator model. We looked carefully at each of these three constructs and at all of the indicators. 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 is provided in Appendix.

After the CFA, four indicators were identified that comprised the curriculum construct (see Appendix): (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).

Five tightly aligned indicators were identified in the ecology construct, which we renamed discourse to better reflect the identified indicators (see Appendix): (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).

We then considered the 12 indicators that were no longer associated with any of the three constructs. First, we completely eliminated four indicators that previously belonged to the ecology construct. Since those working with face validity issues had previously questioned the importance of four indicators, which assessed the physical attributes of the classroom, and since they did not seem to fit the CFA model, we decided to eliminate them from the protocol.

This left eight unmatched indicators. Because we were striving for a parsimonious model, we considered omitting these eight indicators. However, a fourth construct, assessment, with five indicators emerged from the remaining indicators (see Appendix): (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).

This left three of the 26 original indicators still unaccounted for (1) *teacher content knowledge*, (2) *meaningful context*, and (3) *fundamental ideas*. Although all three indicators have a perceived value both by the researchers and the literature, we removed these items from the final model. First, the team felt that *teacher content knowledge*, though critical, is a much broader variable than can be fairly assessed within a single observation. Second, *meaningful context* was deleted as an indicator because it was difficult to measure it consistently and because we had considerable disagreement regarding what the indicator meant in the different domains. Finally, we deleted *fundamental ideas* because, without always seeing the lessons previous and subsequent to the observation, we were often unable to determine how well the teacher tied the lesson to key ideas in the discipline.

We also conducted several additional tests to validate the model. Because of the complexity associated with SEM, absolute parameters are difficult to find, but all parameters fell within acceptable commonly reported boundaries. Specifically, χ^2 is significant $p<0.001$, $\chi^2/df \leq 2$ indicates reasonable fit (Kline, 2005), root mean square error of approximation of 0.1 is on the

TABLE 1

Reliability comparison of EQUIP models

Model	*Indicators*	*Mean*	*Variance*	*Chronbach's α*	*Standardized α*	*Cohen's kappa*
Three constructs						
Instruction	9	2.45	0.077	0.882	0.885	0.56
Curriculum	8	2.30	0.016	0.887	0.889	0.56
Ecology[a]	9	2.37	0.112	0.881	0.880	0.55
Four constructs						
Instruction	5	2.51	0.026	0.898	0.900	0.60
Curriculum	4	2.29	0.014	0.858	0.857	0.56
Discourse	5	2.18	0.013	0.912	0.913	0.51
Assessment	5	2.21	0.024	0.820	0.826	0.64

[a]Ecology is renamed to interaction as the final model is developed

threshold of reasonable fit (Browne & Cudeck, 1993), standardized root mean square residual < 0.1 is considered favorable (Kline, 2005), and the computerized fit index of > 0.90 is considered a good fit (Hu & Bentler, 1999). The four-construct model 19-indicator model, though not quite as parsimonious as a 14-indicator model, provides a good-fitting model that also is solidly supported by the literature base regarding effective inquiry instruction. Further, when the α-values and κ scores of the four-construct model are compared to the original model, reliability remains high (see Table 1). Appendix shows all four constructs with their respective indicators along with the level 3 (proficient) descriptive rubric.

To summarize, we took several steps to assess the validity of EQUIP. First, we tested the entire set of 26 indicators mapped to three constructs. This model was trimmed to find a solid data-driven model that contained three constructs with 14 total indicators. Finally, we arrived at a four-construct model that is justified both from the data and from the literature.

TABLE 2

Goodness-of-fit indicators of models for EQUIP constructs (n=102)

Model	*Indicators*	χ^2	*df*	χ^2/df	*CFI*	*RMSEA*	*SRMR*
Three constructs	26	596.55*	296	2.02	0.834	0.100	0.070
Three constructs	14	152.90*	74	2.07	0.932	0.102	0.052
Four constructs	19	294.65*	146	2.02	0.903	0.100	0.067

*$p < 0.001$

Both the trimmed three-construct model and the four-construct model provided a good fitting model (see Table 2).

Discussion and Implications

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. From the outset, EQUIP 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.

Because many protocols are available, it is important to know which one is best suited for your needs. As such, we would like to be clear regarding what distinguishes EQUIP from the other instruments: (1) EQUIP was developed to function both in mathematics and science classrooms, (2) EQUIP was developed formatively with continual refinements being made until the pragmatic implementation aligned with the theoretical underpinnings and supported by statistical analysis, (3) EQUIP was developed to work with a variety of instructional models (e.g., 5E, 4E×2, learning cycle), (4) EQUIP was developed to allow three distinct levels of assessment (individual indicator level, construct level, and entire lesson level), and (5) EQUIP was developed to be beneficial to program reviewers, researchers, and teacher leaders, as well as to practicing teachers.

The strong face and construct validity in addition to the confirmatory factor analysis allow EQUIP to be used to look at the macro and micro issues associated with inquiry instructional practice. Specifically, the rubrics associated with the individual indicators can be explored with teachers to see individual areas where they can refine their instruction, perhaps one indicator at a time. The composite look at each construct allows for a broader conversation regarding the planning for and implementation of inquiry-based instruction. Similarly, a macro view of inquiry instruction emerges when the composites of the four constructs are summarized to provide a holistic view of the lesson relative to inquiry-based instruction. Finally, when EQUIP is used over time, changes in inquiry instruction can highlight transformations that have occurred.

Even though the context defined in this manuscript was for a professional development experience framed by the 4E×2 Instructional Model, the descriptive rubric for each indicator within EQUIP is written so that observations for all science and math classes can be scored on the instrument. With so much emphasis placed on inquiry instruction, we need a tool to assess its quality. EQUIP takes a large step in helping us accomplish exactly that.

ACKNOWLEDGMENTS

We would like to thank Martie Thompson with the Center for Research and Collaborative Activities at Clemson University for the conversations and assistance that she provided regarding the SEM tests used for the confirmatory factor analysis.

Appendix

Sections IV–VII of EQUIP (Marshall et al., 2008)

Indicator measured	*Pre-inquiry (level 1)*	*Developing inquiry (level 2)*	*Proficient inquiry (level 3)*	*Exemplary inquiry (level 4)*
IV. Instructional factors				
I1. 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
I2. 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
I3. Teacher role	Teacher was center of lesson; rarely acted as facilitator	Teacher was center of lesson; occasionally acted as facilitator	Teacher frequently acted as facilitator	Teacher consistently and effectively acted as a facilitator
I4. Student role	Students were consistently passive as learners (taking notes, practicing on their own)	Students were active to a small extent as learners (highly engaged for very brief moments or to a small extent throughout lesson)	Students were active as learners (involved in discussions, investigations, or activities, but not consistently and clearly focused)	Students were consistently and effectively active as learners (highly engaged at multiple points during lesson and clearly focused on the task)

I5. Knowledge acquisition	Student learning focused solely on mastery of facts, information, and/or rote processes	Student learning focused on mastery of facts and process skills without much focus on understanding of content	Student learning required application of concepts and process skills in new situations	Student learning required depth of understanding to be demonstrated relating to content and process skills
V. Discourse factors				
D1. 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
D2. Complexity of questions	Questions focused on one correct answer; typically short answer responses	Questions focused mostly on one correct answer; some open response opportunities	Questions challenged students to explain, reason, and/or justify	Questions required students to explain, reason, and/or justify. Students were expected to critique others' responses
D3. Questioning ecology	Teacher lectured or engaged students in oral questioning that did not lead to discussion	Teacher occasionally attempted to engage students in discussions or investigations but was not successful	Teacher successfully engaged students in open-ended questions, discussions, and/or investigations	Teacher consistently and effectively engaged students in open-ended questions, discussions, investigations, and/or reflections
D4. Communication pattern	Communication was controlled and directed by teacher and followed a didactic pattern	Communication was typically controlled and directed by teacher with occasional input from other students; mostly didactic pattern	Communication was often conversational with some student questions guiding the discussion	Communication was consistently conversational with student questions often guiding the discussion

D5. Classroom interactions	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
VI. Assessment factors				
A1. 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
A2. 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
A3. Student reflection	Teacher did not explicitly encourage students to reflect on their own learning	Teacher explicitly encouraged students to reflect on their learning but only at a minimal knowledge level	Teacher explicitly encouraged students to reflect on their learning at an understanding level	Teacher consistently encouraged students to reflect on their learning at multiple times throughout the lesson; encouraged students to think at higher levels
A4. Assessment type	Formal and informal assessments measured only factual, discrete knowledge	Formal and informal assessments measured mostly factual, discrete knowledge	Formal and informal assessments used both factual, discrete knowledge and authentic measures	Formal and informal assessment methods consistently and effectively used authentic measures

A5. Role of assessing	Teacher solicited predetermined answers from students requiring little explanation or justification	Teacher solicited information from students to assess understanding	Teacher solicited explanations from students to assess understanding and then adjusted instruction accordingly	Teacher frequently and effectively assessed student understanding and adjusted instruction accordingly; challenged evidence and claims made; encouraged curiosity and openness
VII. Curriculum factors				
C1. Content depth	Lesson provided only superficial coverage of content	Lesson provided some depth of content but with no connections made to the big picture	Lesson provided depth of content with some significant connection to the big picture	Lesson provided depth of content with significant, clear, and explicit connections made to the big picture
C2. Learner centrality	Lesson did not engage learner in activities or investigations	Lesson provided prescribed activities with anticipated results	Lesson allowed for some flexibility during investigation for student-designed exploration	Lesson provided flexibility for students to design and carry out their own investigations
C3. Integration of content and 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
C4. Organizing and 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 nonprescriptive ways	Students organized and recorded information in nonprescriptive ways that allowed them to effectively communicate their learning

REFERENCES

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

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

American Association for the Advancement of Science. (1998). *Blueprints for reform*. New York: Oxford University Press.

Atkin, J., & Karplus, R. (1962). Discovery of invention? *Science Teacher, 29*(5), 45.

Banilower, E. R. (2005). A study of the predictive validity of the LSC Classroom Observation Protocol [electronic version]. Retrieved October 17, 2008, from http://www.horizon-research.com/reports/2005/COP_validity.phprl.

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.

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

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

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.

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

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

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

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.

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: 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.

Chinn, C. A., & Brewer, W. F. (1998). Theories of knowledge acquisition. In B. J. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 97–113). Great Britain: Kluwer Academic.

Cobb, P., Wood, T., & Yackel, E. (1990). Classrooms as learning environments for teachers and researchers. In R. B. Davis, C. A. Maher & N. Noddings (Eds.), *Constructivist views of the teaching and learning of mathematics* (pp. 125–146). Reston: NCTM.

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

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

Eisenkraft, A. (2003). Expanding the 5E model: A proposed 7E model emphasizes “transfer of learning” and the importance of eliciting prior understanding. *The Science Teacher, 70*(6), 56–59.

Henry, M., Murray, K. S., & Phillips, K. A. (2007). *Meeting the challenge of STEM classroom observation in evaluating teacher development projects: A comparison of two widely used instruments*. St. Louis: Henry Consulting. Document Number.

Horizon Research. (2002). Inside the classroom interview protocol [electronic version]. Retrieved May 14, 2008, 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.

Interstate New Teacher Assessment and Support Consortium (INTASC). (1992). *Model standards for beginning teacher licensing and development: A resource for state dialogue*. Washington, DC: Council for Chief State School Officers. Retrieved December 13, 2002.

Karplus, R. (1977). Science teaching and the development of reasoning. *Journal of Research in Science Teaching, 14*, 169.

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

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

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

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

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.

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

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

Llewellyn, D. (2005). *Teaching high school science through inquiry: a case study approach*. Thousand Oaks: Corwin.

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

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

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.

Marshall, J. C., Horton, B., & Smart, J. (2009). 4E×2 Instructional Model: Uniting three learning constructs to improve praxis in science and mathematics classrooms. *Journal of Science Teacher Education* (in press).

Marshall, J. C., Horton, B., Smart, J., & Llewellyn, D. (2008). EQUIP: Electronic Quality of Inquiry Protocol [electronic version]. Retrieved May 30, 2008, from www.clemson.edu/iim.

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

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

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

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

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

Moscovici, H., & Holdlund-Nelson, T. (1998). Shifting from activity mania to inquiry. *Science and Children, 35*(4), 14–17.

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

National Board for Professional Teaching Standards. (2006). Making a difference in quality teaching and student achievement. Retrieved October 23, 2006, from http://www.nbpts.org/resources/research.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston: NCTM.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston: NCTM.

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

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

Piburn, M., & Sawada, D. (2001). Reformed Teaching Observation Protocol (RTOP): Reference manual [electronic version]. *ACEPT Technical Report No. IN00-3*. Retrieved Oct. 17, 2008, from http://physicsed.buffalostate.edu/AZTEC/RTOP/RTOP_full/PDF/.

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.

Saam, J., Boone, W. J., & Chase, V. (2000). A snapshot of upper elementary and middle school science teachers' self-efficacy and outcome expectancy [Electronic Version]. Retrieved June 15, 2009, from www.eric.ed.gov.

Sampson, V. (2004). The Science Management Observation Protocol. *The Science Teacher, 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)*: Arizona State University. Document Number.

Schmidt, W. H., McNight, C. C., & Raizen, S. A. (2002). A splintered vision: An investigation of U.S. science and mathematics education. 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: The Free.

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.

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

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

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

Eugene T. Moore School of Education
Clemson University,
418G Tillman Hall, Clemson, SC 29634-0705, USA
E-mail: marsha9@clemson.edu