THE CREATION, VALIDATION, AND RELIABILITY ASSOCIATED WITH THE EQUIP (ELECTRONIC QUALITY OF INQUIRY PROTOCOL): A MEASURE OF INQUIRY-BASED INSTRUCTION

Abstract

K-12 science teachers self-report that 39% of classroom instructional time is devoted to inquiry-based instructional practice, but the quality of this instruction is largely unknown. Current observational protocols seem either too broad by looking at instructional practice in its entirety (e.g., classroom management, instructional practice, assessment), or they seem too specific by considering only one aspect of inquiry-based instruction. Therefore, there was a perceived need to develop a protocol that seeks to look at the major constructs of inquiry-based instructional practice. The resulting protocol assesses 19 indicators aligned with four constructs: Instruction, Curriculum, Assessment, and Interactions. 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.

Jeff C. Marshall, Clemson University

K-12 mathematics and science teachers report on average that 39% of classroom instructional time is devoted to inquiry-based instructional practice (Marshall, Horton, Igo, & Switzer, In Press). These data provide an understanding of teacher perception, but they do not provide a standardized way to examine the quality of the inquiry-based instruction that is being led. Since perceptions and definitions of inquiry-based instruction vary widely (Anderson, 2002), it is important to be able to clarify what is meant by inquiry-based instruction and provide a solid measurement of the components entailed in this inquiry-based teaching and learning.

Informed by insights from multiple educational theories and philosophies (Bransford, Brown, & Cocking, 1999; Dewey, 1938; National Research Council, 1996), the following definition of inquiry-based instruction will be adopted for this paper: "A student-centered pedagogy that uses purposeful, extended investigations set in the context of real-life problems as both a means for increasing student capacities and as a feedback loop for increasing teachers' insights into student thought processes" (Supovitz, Mayer, & Kahle, 2000, p. 332). By operationalizing inquiry-based instruction, we have provided a foundation to allow us to explore the creation, the validity, and the reliability associated with a protocol, EQUIP (Electronic Quality of Inquiry Protocol), that seeks to measure it.

Since our intent was to measure the quality of inquiry-based instruction that was occurring in the classroom, our needs were only partially addressed by any one of these instruments. Therefore, a new protocol was developed that is informed by multiple existing frameworks (Horizon Research, 2002; Llewellyn, 2007; Sampson, 2004; Sawada et al., 2000). A new protocol was necessary, instead of cropping from multiple instruments, so that a coherent framework could be

assembled that allows reliability and validity issues to be addressed consistently (Henry, Murray, & Phillips, 2007). The Electronic Quality of Inquiry Protocol (EQUIP) was designed to provide a framework for evaluating the quality of inquiry-based instruction in science and math classrooms. The instrument was developed around the following instructional factors: usage of time, instruction, curriculum, and ecology/climate of the classroom. Merely encouraging teachers to implement inquiry-based practices is not sufficient. Teacher educators need a tool to assess the quality of inquiry if they are to promote and support inquiry-based instruction in the classroom.

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)

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.

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 et al., 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. We ultimately 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.

The *Science Teacher Inquiry Rubric (STIR)* (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.

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 Electronic Quality of Inquiry Protocol (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 et al., 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 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

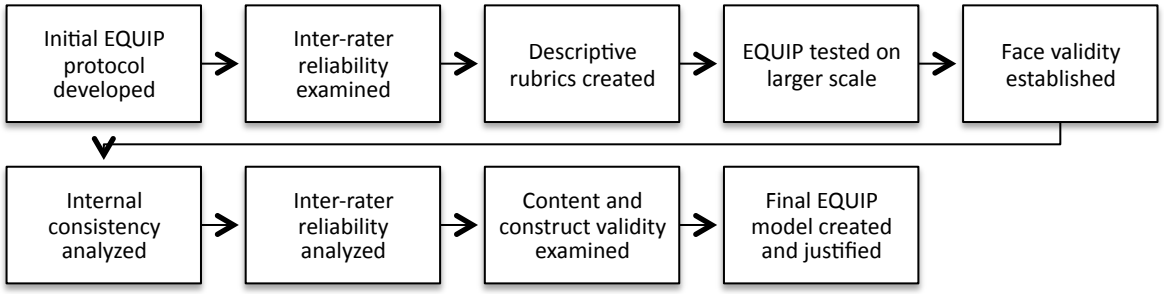
Context of Development

As part of a professional development 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 professional development 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 x 2 Instructional Model (Marshall, Horton, & Smart, In Press). We know that 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 x 2 Instructional Model integrates these learning constructs into a single dynamic model that is used to guide transformation of instructional practice.

The 4E x 2 Instructional Model builds upon the 5E Instructional Model (Bybee et al., 2006) and other similar models (Atkin & Karplus, 1962; Bybee et al., 2006; Eisenkraft, 2003; Karplus, 1977) that focus on inquiry instruction. By integrating inquiry instruction with formative assessment and teacher reflection, a single, cohesive model is formed. To guide and assess teachers' transformation to inquiry-based instruction using the 4E x 2, we undertook the challenge of developing and validating EQUIP, outlined in Figure 1. However, we designed EQUIP broadly enough to measure inquiry instruction that does not align with the 4E x 2.

Figure 1: Flowchart of the design and validation of EQUIP



Development: Semester One

Initial EQUIP protocol. The development of EQUIP began with two primary steps: (1) drawing 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. From the literature, we identified the following initial categories that guided early forms 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. The first version of the protocol was heavily influenced by the RTOP and the Inside the Classroom Observational Protocol. In addition to some of the initial categories, these existing protocols provided a framework for initial development of a format to assess use of instructional time, form of assessments, and grouping of items.

Inter-rater reliability. We piloted the initial version of EQUIP in high school science and math classrooms for one academic semester. Our team of three researchers, a science educator, a math educator, and a curriculum and instruction doctoral student, conducted individual and paired observations in order to assess inter-rater 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. Additionally, by developing a descriptive rubric, raters would become more systematic and less subjective during observations, thereby bolstering instrument reliability. Finally, we decided to create a descriptive rubric 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.

During this 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, we had a form of EQUIP that was ready for more rigorous testing. This form contained seven discrete sections. 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 were assessed by a total of 26 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 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 five 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 following sections.

Instrument Validation

Research Team and Observations

With the addition of another doctoral student in Curriculum and Instruction, our research team now 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 pre-recorded 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 five months. All observations were in middle school math and science classrooms. All data was entered into Microsoft Access, converted into an Excel spreadsheet, and then used SPSS and Mplus for analysis.

Validity

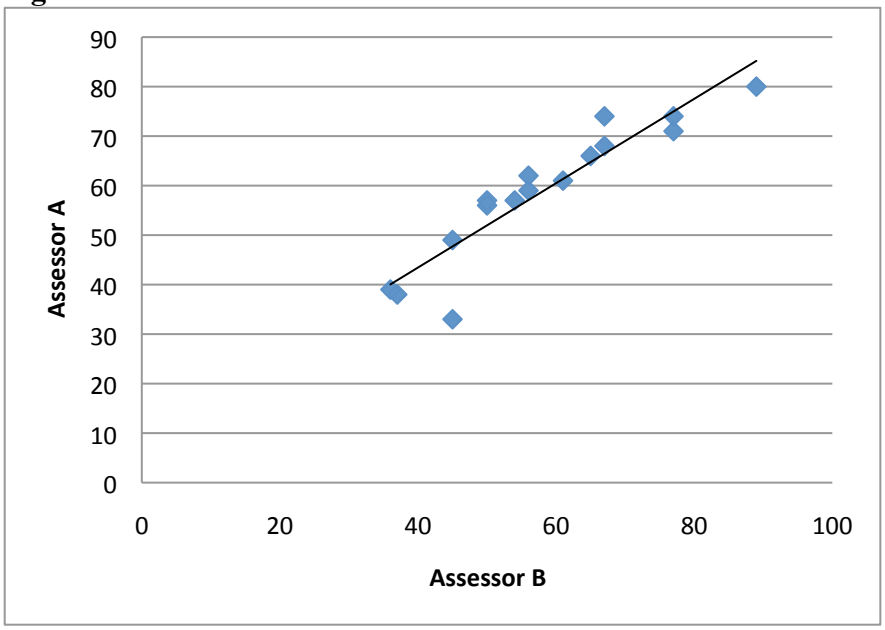
Face validity. In addition to the team members, four science and three math education researchers who were familiar with the underlying constructs being measured by the instrument helped assess the face validity. Further, two measurement experts with knowledge of instrument development assessed the instrument structure. To establish face validity, we posed the following questions: Does EQUIP seem like a reasonable way to assess the quality of inquiry? Does it seem well designed? Does it seem as though it will work reliably? For the content specialists, we had one more question: Does it maintain fidelity to the discipline (math/science)? Their responses assured us that EQUIP did indeed possess face validity.

Internal consistency. EQUIP indicators were examined for internal consistency using Cronbach's Alpha (α) for all 102 class observations. The α-value ranged from .880-.889, demonstrating strong internal consistency. For the science observations ($n = 60$), the standardized α-value ranged from .869-.874, and for the math observations ($n = 42$), the range was .823-.861. Thus, the instrument items hold together well as a whole, and for science and mathematics separately.

Inter-rater reliability. We conducted 16 paired observations to analyze inter-rater reliability, via Cohen's Kappa (κ). The κ scores averaged .61 for the nine indicators for Instruction, .62 for the eight indicators for Curriculum, and .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 .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 visa versa. This value was generated using a summative score that included all 26 indicators plus the 5 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 .958 and .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 model 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 .858-.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 were identified that were tied to Instruction: (1) *instructional strategies*, (2) *order of instruction*, (3) *teacher role*, (4) *student role*, and (5) *knowledge acquisition*. The literature base to support the content validity associated with these Instruction indicators include the following works: Abell & Lederman (2007); Bransford et al. (2000); Bybee et al. (2006); Chiappetta & Koballa (2006); Mortimer & Scott (2003); and NCR (2000).

After the CFA, four indicators were identified that comprised the Curriculum construct: (1) *content depth*, (2) *learner centrality*, (3) *standards*, and (4) *organizing and recording information*. Literature to support the Curriculum construct indicators includes the following: Donovan & Bransford (2005); Knowles & Brown (2000); Llewellyn (2002, 2007); Luft, Bell, & Gess-Newsome (2008); Marzano, Pickering, & Pollock (2001); NBPTS (2000); NRC (1996); Schmidt, McNight, & Raizen (2002); and Wiggins & McTighe (1998).

Five tightly aligned indicators were identified in the Ecology construct, which we renamed Discourse to better reflect the identified indicators: (1) *questioning level*, (2) *complexity of questions*, (3) *questioning ecology*, (4) *communication pattern,* and (5) *classroom interaction.* A sample of the literature base to support the content validity for the five identified indicators include: Ball & Cohen (1999); Chin (2007); Kelly (2007); Lampert (1990); Lemke (1990); Moje (1995); Morge (1995); and van Zee, Iwasyk, Kursoe, Simpson, & Wild (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 team members had previously questioned the importance of four indicators, which assessed the physical attributes of the classroom, and since they didn't 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: (1) *prior knowledge*, (2) *conceptual development*, (3) *student reflection*, (4) *assessment type(s)*, and (5) *role of assessing*. The rationale to include Assessment as a construct of effective inquiry instruction is supported by several works, including: Black & Wiliam (1998); Bransford et al. (2000); Driver, Squires, Rushworth, & Wood-Robinson (1994); Stigler & Hiebert (1999).

This left three indicators 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 $p < .001$, $\chi^2/df \leq 2$ 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, 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 Figure 3). The entire descriptive rubric that details all four constructs with their respective indicators can be found at www.clemson.edu/iim.

Figure 3: Reliability comparison of EQUIP models

Model	Indicators	Mean	Variance	*Chronbach α*	*Standardized α*	*Cohen's Kappa*
Three constructs						
Instruction	9	2.45	.077	.882	.885	.56
Curriculum	8	2.30	.016	.887	.889	.56
Ecology*	9	2.37	.112	.881	.880	.55
Four constructs						
Instruction	5	2.51	.026	.898	.900	.60
Curriculum	4	2.29	.014	.858	.857	.56
Discourse	5	2.18	.013	.912	.913	.51
Assessment	5	2.21	.024	.820	.826	.64

*Ecology was renamed to Discourse as the final model was developed.

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. Both the trimmed three-construct model and the four-construct model provided a good fitting model (see Figure 4).

Figure 4: 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	.834	.100	.070
Three constructs	14	152.90***	74	2.07	.932	.102	.052
Four constructs	19	294.65***	146	2.02	.903	.100	.067

***p < .001.

Discussion and Limitations

Because of the complex 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. The culminating four-construct (Instruction, Curriculum, Interaction, and Assessment) EQUIP is a reliable and valid instrument that meets these goals.

The EQUIP provides a venue 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 x 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. We believe EQUIP takes a large step in helping us accomplish exactly that.

References

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

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-12.

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

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Skyes

(Ed.), *Teaching as a Learning Profession: Handbook of Policy and Practice*. San Francisco: Jossey-Bass.

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

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

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

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

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, CA: 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)

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.

Dewey, J. (1938). *Experience and education*. New York: Collier Books.

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: Taylor & Francis Ltd.

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*: M.A. Henry Consulting, LLCo. 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.

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, NJ: Lawrence Erlbaum Associates.

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

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

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, NJ: Ablex.

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

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

Llewellyn, D. (2007). *Inquiry within: Implementing inquiry-based science standards in grades 3-8* (2nd 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., Horton, B., Igo, B. L., & Switzer, D. M. (In Press). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom *International Journal of Science and Mathematics Education*.

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

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

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.

Mortimer, E. F., & Scott, P. H. (2003). *Meaning making in secondary science classrooms*. Maidenhead, UK: Open University 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 Evaluationo. 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. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM, Inc.

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.

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.

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)*: Arizona State Universityo. 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

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 Press.

Supovitz, J. A., Mayer, D. P., & Kahle, J. B. (2000). Promoting inquiry-based instructional practice: The longitudinal impact of professional development in the context of systemic reform. *Educational Policy, 14*, 331-356.

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.

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