International Journal of Research & Method in Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cwse20

Measuring the impact of professional development on science teaching: a review of survey, observation and interview protocols

Barbara Heath [a] , Aruna Lakshmanan [a] , Aaron Perlmutter [a] & Lori Davis [a]

[a] East Main Educational Consulting, LLC , Wilmington, NC, USA
Published online: 06 Apr 2010.

International Journal of Research & Method in Education
Vol. 33, No. 1, April 2010, 3–20

Measuring the impact of professional development on science teaching: a review of survey, observation and interview protocols

Barbara Heath*, Aruna Lakshmanan, Aaron Perlmutter and Lori Davis

East Main Educational Consulting, LLC, Wilmington, NC, USA

(*Received 16 December 2008; final version received 10 March 2009*)

Instrument choice is a crucial part of evaluation of professional development programmes. The use of multiple evaluation methods helps in triangulation, and offers insight into the developmental sequence involved in the changes in teacher beliefs and practice. Most current instruments are self-contained and not designed for use in conjunction with other instruments. Linking such instruments poses challenges for evaluators. We performed an extensive literature survey to identify instruments best aligned with each other. We describe here our evaluation needs, examine the subset of instruments selected based upon these needs, and discuss our choice of instruments guided by specific principles.

Keywords: educational measurement; general methodology; mixed methods; review; survey research

Introduction

A range of tools designed to measure the impact of professional development on practicing K-12 teachers have been developed over the last few decades. These include surveys, classroom observation protocols, and interviews, most of which have been developed as stand alone measures. However, a comprehensive evaluation of a professional development programme often requires the usage of a set of instruments to be used in conjunction with one another. In this article, we describe the results of an extensive review of instruments and examine the factors that led us to choose a set of instruments for our use.

The emergence of systemic reform as a major educational policy initiative in the 1990s resulted in teacher professional development being given a new prominence. The National Education Goals Panel (1999), which set targets for standards-based education reform, designated 'teacher education and professional development' as one of the goals. By the late 1990s, more than 25 states had enacted legislation that called for improved teacher professional development (Darling-Hammond 1997), the rationale being that teacher professional development plays an important role in improving student achievement. In fact, it has been suggested that standards-based reform is 'doomed to failure' unless interventions that provide teachers with the skills, knowledge and attitudes required to use new strategies are developed (Duttweiler and McEvoy 1999).

*Corresponding author. Email: bheath@emeconline.com

ISSN 1743-727X print/ISSN 1743-7288 online

DOI: 10.1080/17437270902947304
http://www.informaworld.com

The literature reports the difficulty of defining a direct relationship between teacher knowledge and pedagogy to student achievement, and even greater difficulty in links between specific professional development and student achievement (Darling-Hammonds and Sykes 1999; National Research Council 2000). Guskey's (2000) levels of professional development evaluation shows student learning outcomes as the most complex and requiring success at earlier levels such as teachers' learning and teachers' use of new knowledge and skills. Determining the impact on teachers represents a necessary step prior to determining if the professional development may be contributing to student learning. In this regard, our efforts focus first on determining observable impacts on teachers prior to attempting linkages to student achievement, thereby attaining level four (participants' use of new knowledge and skills) of Guskey's evaluation levels.

Garet and others (2001), using a large national sample of mathematics and science teachers, studied the effects of various characteristics of professional development programmes on teachers' knowledge, skills, and classroom practice. They concluded that effective professional development programmes (those that positively change teachers knowledge, skills, and/or classroom practice), should: (1) focus on content – what teachers learn during professional development – rather than on the process and delivery system; (2) offer opportunities for active (hands-on) learning; and (3) be integrated with the daily life of the school (i.e. coherence). Longer and more intensive programmes were in general found to be more effective than programmes with fewer contact hours. Kennedy (1998) also emphasizes the ineffectiveness of the 'one-shot workshop'. Further research confirms that short-term workshops are one of the least effective ways of training adults (Birman et al. 2000).

Additionally, Garet et al. (2001, 934) found that enhanced teacher knowledge and skills have a 'substantial positive impact on change in teacher practice'. Teachers who reported improvement in knowledge and skills were more likely to report changing their teaching practices. Teachers who experienced coherent professional development were more likely to change their teaching practice, indicating that building on what teachers have already learned, aligning professional development with state and district standards and assessment, and encouraging better communication among teachers with similar ideas all impact instructional practice (Resnick 2005).

Context

The importance of professional development for teachers has led to the development of several initiatives to enhance teacher learning. One such initiative is the mathematics and science partnerships (MSP) programme funded by the Department of Education through formula grants awarded to individual states. The MSP programme is designed to improve teacher content knowledge and student achievement by encouraging: improvement of mathematics and science teacher education at the Institutes of Higher Education (IHEs); focus on education of mathematics and science educators as a career-long process; collaboration between teachers and scientists, mathematicians, and engineers; and ongoing professional development for teachers to improve their knowledge and pedagogical skills (US Department of Education 2008).

In 2006, the state of North Carolina awarded an MSP grant to a county located in the southeastern region of the state. The MSP project, now in its third and final year, aims to increase the content knowledge and improve the pedagogy of science and math teachers in grades five through eight. The project seeks to enhance student achievement through its efforts with teachers. Teachers participate in science professional development in the form of content courses and professional learning groups. The professional development offered to teachers in this project aligns with the criteria suggested by Garet et al. (2001) in the above section.

The authors of this article were contracted as the external evaluators for the project. Our main evaluation goals were: (1) to determine if the knowledge and skills learned are applied by teachers in the classroom, and (2) to gain a better understanding of teacher beliefs and factors that influence the extent of reform implementation.

However, professional development is an ongoing and systematic process, and change does not happen overnight (Glusac 2008). Often, change in beliefs precedes change in practices (Bandura 1977) and frequently teachers are able to talk about changing their practice prior to actually demonstrating change in their classroom (National Research Council 2000). Based on this developmental sequence, we decided to study teacher beliefs with respect to their ability to shift toward a student-centred classroom, often referred to as reformed teaching, by using a survey to measure changes in self-efficacy, followed by classroom observations to measure the extent of reform practice, and lastly teacher interviews to gain more in-depth knowledge of the issues involved in the application of reform teaching.

Initially, we adopted instruments that met our data collection needs but were not developed in concert, including the *Science Teaching Efficacy Belief Instrument-A* (STEBI-A) (Riggs and Enochs 1990), *Reformed Teacher Observation Protocol* (RTOP) (Piburn and Sawada 2000), and an in-house developed interview protocol. Using multiple methods of data collection helps to overcome some of the weaknesses associated with a particular method and provides a more complete understanding of the phenomenon under study. Using multiple tools also helps with data and methodological triangulation, which are recommended to test for consistency and minimize errors associated with any one particular method (Lincoln and Guba 1985; Patton 2002).

However, we experienced the difficulties associated with linking stand-alone instruments to achieve a multiple methods evaluation. The issue stems from the fact that there is not a set of consistent and complementary instruments (survey/observation/interview) in the science education literature, but rather many individual instruments designed to measure different classroom outcomes. While observation/teacher interview protocols do exist, the interviews in these protocols are often designed to complement the observation parts rather than examine issues related to the implementation of inquiry-based teaching in the science classroom (Lawrenz, Huffman, and Appledoorn 2002; Wainwright, Flick, and Morrell 2003; Weiss et al. 2003). Thus, the selection of instruments requires a careful and time consuming examination of the items therein to determine if the items are addressing similar themes.

A range of surveys, observations, and interview protocols have been developed and applied to evaluate the impact of professional development over the past two decades. However, many of these tools are stand-alone and have not been developed with triangulation or measurement of validity in mind. In this article, we outline our needs as we prepare for additional MSP programme evaluations, examine some of the more promising instruments selected as a result of an extensive survey of the literature and discuss our choice of instruments based on specific guiding principles.

Methods

The instrument collection assembled for this literature review is based on references provided by experts in the field of science and mathematics education and evaluation along with additional literature searches through databases such as the Education Resources Information Center (ERIC) database. Keywords used in the database searches included combinations of the following: science, teacher development, evaluation, reformed teaching, inquiry, performance (classroom), measures, and professional development. From the two sources, an expansive list of documents was created.

Additional selection criteria were instituted which included these key factors:

(1) Focus on science
(2) Survey, interview, or observation instrument
(3) Limited to teachers as the data source, not students
(4) Grade-level independent
(5) Statistical validity and reliability data were available

Applying the selection criteria resulted in the review of numerous articles and their associated instruments. The articles were assembled in a binder and assigned to one of five reviewers. Reviewers were responsible for reading and presenting the information from the article to the reviewer group. Each instrument was discussed and its information noted in a table for referral purposes.

Results

Based on the review described above, 22 instruments were selected for further examination. These included 11 surveys, five observation protocols, and six interview protocols. This subset of instruments was evaluated in details. For each instrument, we examined the grade levels, number of items, statistical properties, measurement scales, outcomes, key dimensions, and theoretical basis. The results are presented below in Table 1 (surveys), Table 2 (observation protocols), and Table 3 (interview protocols), along with our additional commentary.

Surveys

A very broad range of teacher surveys are available, as reflected in the 11 instruments selected for this review. These surveys vary in scope, and were selected as they offer many different starting points for a researcher studying reform teaching. Many of those selected are designed specifically to measure teacher beliefs, attitudes and/or self-efficacy, as there is evidence in teacher development literature that these affects are related to teacher classroom practice (Posnanski 2002; Khourey-Bowers and Simonis 2004; Crippen 2008). Of the 11 surveys examined, two are concerned with teacher self-efficacy – *Science Teaching Efficacy Belief Instrument-Preservice* (STEBI-B), and the *Teaching of Science as Inquiry* (TSI), two measured other teacher beliefs such as beliefs relating to science teaching environments – *Context Beliefs about Teaching Science* (CBATS), and beliefs about teaching and learning science *Beliefs about Reformed Science Teaching and Learning* (BARSTL), and one measured teachers' affective states as they enter professional development activities – *Pedagogical*

Discontentment Questionnaire (PDQ). Other instruments examined teachers' perceptions on science, scientific inquiry, and their understanding of reform and constructivist methods – *Scientific Attitude Inventory – II* (SAI-II), *Views of Nature of Science* (VNOS), *Teaching, Learning and Computing*, *Teaching and Learning about Inquiry*, and the *Thinking about Science Survey* (TSSI), while one – *Constructivist Learning Environment Survey* (CLES) – examined teacher perceptions of the classroom learning environment.

Many of the survey instruments studied were suited for K-12 grades, while a few were restricted to elementary or secondary grades. The development of some of the instruments is based on theory, but others seem to have stemmed from more pragmatic needs. They also vary in the extent and methods used to establish reliability and validity. For example, the SAI-II establishes reliability through high calculations on: (1) the split-half reliability coefficient; and (2) Cronbach's alpha reliability coefficient (Moore and Foy 1997). Face validity is claimed based on the determinations of a panel of judges chosen to rate alignment of attitude and position statements from the original SAI (Moore and Foy 1997). However, construct validity could not be established with the SAI-II. The authors present possible explanations for this finding, but an overall determination of the instrument's validity is complicated by the lack of validity support from the factor analysis. On the other hand, the authors had not performed a factor analysis, and it would appear that there were no issues with the instrument's validity, because it would not be subject to the same level of rigorous testing. Examples such as this precluded a simple yes/no determination of an instrument's reliability and validity in Tables 1–3.

Classroom observation protocols

In evaluating professional development, it is important to assess the observable impact that it has on teachers'content and teaching skills. Requirements of federally funded research initiatives such as the MSPs have led to a number of classroom observation protocols being developed in recent years (Dirr 2002). Emphasis on inquiry and conceptual understanding is a key component of reformed teaching, and the extent to which a teacher incorporates these in his/her lesson can be best captured using classroom observation protocols. Classroom observation protocols use a range of approaches including checklists, narrative descriptions, questionnaires, and classifications. For this review, we focused on instruments that have been well researched and developed using rigorous methods. We investigated five protocols including instruments that are primarily quantitative as well as instruments that have both quantitative and qualitative components. Two of the five instruments, the *Inside the Classroom – Classroom Observation Protocol* (ITC-COP) and the RTOP, were developed first and are widely used. Two others, the *Oregon Teacher Observation Protocol* (OTOP) and the *Collaboratives for Excellence in Teacher Preparation – Classroom Observation Protocol* (CETP-COP), used the RTOP and the ITC-COP (or its antecedents) in their development. Three of these (ITC-COP, CETP-COP, and the OTOP) accompany a teacher interview protocol, *Inside the Classroom – Teacher Interview Protocol* (ITC-TIP), the *Collaboratives for Excellence in Teacher Preparation* teacher interview protocol, and the *Oregon Teacher Interview Protocol* (OTIP), respectively to be administered prior to or after the classroom observation. While the ITC-COP and CETP-COP require an overall or 'capsule' rating, the RTOP does not. It has an overall numeric score for each participant, while the OTOP

Table 1. Survey instruments.

Instrument (author, date)	Description	Outcome(s)	Comments
Teaching and learning about inquiry (Wee et al. 2007)	Grade level: unspecified # Items: 16 Scale: open-ended and rubric style (more to less orientation toward inquiry)	Teacher understanding of inquiry impact of professional development programs on teacher inquiry	Qualitative design Multiple data sources collected throughout professional development Small sample size
Pedagogical discontentment survey (Southerland et al. 2006)	Grade level: K-12 # Items: 30 Scale: no discontentment – very high discontentment	Science teachers' 'affective states' as they enter professional development activities	Models of conceptual change (Strike and Posner 1992; Dole and Sinatra 1998; Feldman 2000; Gregoire 2003) Quantifies teachers' dissatisfaction resulting from the misalignment of their pedagogical beliefs with their instructional practices
TSI (Smolleck, Zembal-Saul, and Yoder 2006)	Grade level: elementary # Items: 69 Scale: strongly disagree – strongly agree	Teacher self efficacy for inquiry-based classrooms	Bandura's social learning theory Intended for pre-service and beginning teachers Investigates the connection between teacher beliefs and 'the teaching of science as inquiry' Extensive instrument development
BARSTL (Sampson and Benton 2006)	Grade level: elementary # Items: 32 Scale: strongly disagree – strongly agree	Teachers' beliefs about teaching and learning science and the alignment of these beliefs with the reform movement in science education	Von Glasersfeld's Constructivism (1989) Subscales: (1) How people learn about science, (2) Lesson design and implementation, and (3) Characteristics of teachers and the learning environment, and the nature of science
VNOS (B) (Abd-El-Khalick, Bell, and Lederman 1998)	Grade level: post-secondary # Items: 6 Scale: open-ended	Factors affecting preservice teacher conceptions of nature of science (NOS) on classroom instruction	Multiple instruments (A–E)[a] Evolution of instruments and methods/procedures Triangulation of data sources: (1) questionnaires, (2) artifacts, including lesson plans, classroom video footage, portfolios, and supervisor observation notes, and (3) interviews

Table 1. (*continued*).

Instrument (author, date)	Description	Outcome(s)	Comments
CBATS (Lumpe, Haney, and Czerniak 2000)	Grade level: K-12 # Items: 26 Scales: strongly disagree – strongly agree; very unlikely – very likely	Beliefs about environmental factors related to science teaching	Ford's motivation systems theory (1992) Bandura's theory of collective efficacy (1997)
TSSI (Cobern 2000)	Grade level: elementary # Items: 60 Scale: strongly disagree – strongly agree	Perception of science based on cultural and social issues	Nine categories quantitatively address cultural and social issues related to science: (1) Epistemology, (2) Science and the Economy, (3) Science and the Environment, (4) Public Policy and Science, (5) Science and Public Health, (6) Science, Religion and Morality, (7) Science, Emotions, and Aesthetics, (8) Science, Race, and Gender, and (9) Science for All
Teaching, learning, and computing: 1998 (Ravitz, Becker, and Wong 2000)	Grade level: 4–12 # Items: 175 Scale: various Likert items	Two overarching teaching approaches: traditional transmission and constructivist-compatible instruction	Constructivist-compatible instruction encompassed the following areas: (1) projects, (2) group work, (3) problem solving tasks, (4) written reflective thought, and other tasks requiring meaningful thought Three types of schools sampled: (1) national probability sample, (2) purposefully selected sample of schools with high measures of computer technology (per capita), and (3) purposefully selected sample of schools involved in educational reform efforts
SAI II (Moore and Foy 1997)	Grade level: 6–12 # Items: 40 Scale: disagree strongly – agree strongly	Attitudes (emotional and intellectual) toward science	5-point Likert scale Six positive and six negative statements comprise 12 position statements Removes gender-based and complicated language from SAI, and reduces number of items from 60

Table 1. (*continued*).

Instrument (author, date)	Description	Outcome(s)	Comments
CLES (Taylor, Fraser, and Fisher 1997)	Grade level: secondary # Items: 30 Scale: never – very often	Perception of classroom learning environment	Student and teacher versions available Based on constructivist theory; measures the extent to which constructivist ideals are implemented Assesses five dimensions of the learning environment: (1) personal relevance, (2) uncertainty, (3) shared control, (4) student negotiation, and (5) critical voice. Available in actual and preferred forms
STEBI-B (Riggs and Enochs 1990)	Grade level: elementary # Items: 23 Scale: strongly disagree – strongly agree	Preservice teachers' science teaching efficacy beliefs	Bandura's social learning theory (1977) Subscales: Personal Science Teaching Efficacy Belief (PSTE) and Science Teaching Outcome Expectancy (STOE)

[a]Lederman and O'Malley (1990); Abd-El-Khalick (1998); Bell, Lederman, and Abd-El-Khalick (2000); Lederman et al. (2001); Bell and Lederman (2003); Bell et al.(2003).

Table 2. Observation protocols.

Instrument (author, date)	Description	Outcome(s)	Comments
STIR (Beerer and Bodzin 2003)	Grade level: K-5 # Items: 6 Scale: teacher-centred – learner-centred	Implementation of inquiry into classroom instruction	Bodzin and Cates web-based inquiry for learning science (WBI) (2002) used to develop STIR Originally created as a self-assessment tool but better used as an observation instrument Each point on the scale is a statement that describes specific teacher and learner behaviours to determine a lesson's alignment with inquiry
OTOP* (Wainwright, Flick, and Morrell 2003)	Grade level: K-16 # Items: 10 Scale: characterizes lesson – not observed	Impact of reform-based professional development on faculty and teachers	Developed with OTIP instrument through OCEPT (NSF-funded collaboration) Designed as a descriptive tool to 'generate a profile of what was happening across institutional settings rather than to assign a score to a particular lesson' Built based on the RTOP
ITC COP* (Weiss et al. 2003)	Grade level: K-12 # Items: 67 Scale: not at all – to a great extent; other scales; open-ended	Degree of reformed teaching based on four domains: (1) design, (2) implementation, (3) content, and (4) classroom culture	Used in combination with pre and post interview; combination of checklists and 5-point scales; requires providing an overall rating and a 'quality capsule' Ratings of 4 categories: (1) design, (2) implementation, (3) mathematics/science content, and (4) classroom culture
CETP-COP (Lawrenz, Huffman, and Appledoorn 2002)	Grade level: K-16 # Items: 12 Scale: not at all – to a great extent	Describing and rating classroom activities using standards-based criteria	Developed as part of the Collaboratives for Excellence in Teacher Preparation (CETP) Program Core Evaluation Project Employs five-minute timed intervals to rate instruction type, student engagement and cognitive activity; other key indicators also rated Requires a 'capsule description' of the overall quality of the lesson Used the RTOP and COP among other instruments in its development

Table 2. (*continued*).

Instrument (author, date)	Description	Outcome(s)	Comments
RTOP* (Piburn and Sawada 2000)	Grade level: K-20 # Items: 39 Scale: never occurred – very descriptive	Reformed classroom instruction	Describes degree of reformed teaching based on: (1) Lesson Design and Implementation, (2) Content: Propositional Knowledge, (3) Content: Procedural Knowledge, (4) Classroom Culture: Communicative Interactions, and (5) Classroom Culture: Student/Teacher Relationships Can be used by anyone who has undergone training

Note: All instruments are constructed to analyse the subject of science; an * denotes that the instrument also includes the subject of math.

Table 3. Interview protocols.

Instrument (author, date)	Description	Outcome(s)	Comments
TBI (Luft and Roehrig 2007)	Grade level: secondary # Items: 7 Scale: open-ended	Science teachers' beliefs	Intended for use with teachers possessing different experience levels Aimed at understanding how beliefs become modified from the pre-service period to the later years of a teaching career Responses to each item can be classified into one of five categories: (1) traditional, (2) instructive, (3) transitional, (4) responsive, and (5) reform-based, thus allowing a profile of beliefs to be generated
Beliefs questions (Luft 2005)	Grade level: unknown # Items: 7 Scale: N/A	Teachers' beliefs about their teaching and their students' learning	Each interview question covers one of the following: (1) the environment, (2) student knowledge, (3) understanding, (4) attention to students & standards, (5) assessment, (6) learning, and (7) student response
OTIP* (Wainwright, Flick, and Morrell 2003)	Grade level: K-16 # Items: 4 Scale: open-ended	Impact of reform-based professional development on faculty and teachers	Developed with OTOP instrument through OCEPT (NSF-funded collaboration) Questions encourage discussion about items within OTOP
ITC TIP* (Weiss et al. 2003)	Grade level: K-12 # Items: 18 Scale: open-ended	Teacher perceptions of influential factors in the selection of lesson content and pedagogy	Intended for use with Inside the Classroom Observation and Analytic Protocol (results are used to complete the second half of it) 18 items address five areas: (1) learning goals, (2) content/topic, (3) resources used to design the lesson, (4) the teacher, and (5) context Some items include probes
VNOS (B) (Abd-El-Khalick, Bell, and Lederman 1998)	Grade level: post-secondary # Items: 10 Scale: open-ended	Factors affecting preservice teacher conceptions of nature of science (NOS) on classroom instruction	Validation of VNOS (B) responses Generation of in-depth participant profiles Triangulation of data sources: (1) questionnaires, (2) artifacts, including lesson plans, classroom video footage, portfolios, and supervisor observation notes, and (3) interviews
LoU (Hall and Hord 1987)	Grade level: K-16 # Items: varies Scale: open-ended	Extent of mastery of an innovation (such as reformed teaching)	Three levels of non-use (non-use, orientation, preparation) and five levels of use (mechanical, routine, refinement, integration, renewal) Branching interview is a quick assessment; focused interview needs a certified interviewer

Note: All instruments are constructed to analyse the subject of science; a* denotes that the instrument also includes the subject of math.

generates a profile rather than assigns a numeric score. The fifth instrument examined, *Science Teacher Inquiry Rubric* (STIR), is a rubric that uses explicit descriptions of behaviour rather than a numeric scale for rating. This instrument can be used in a variety of ways such as a continuing teacher self-assessment tool or periodically by an observer.

Interview protocols

While surveys and observation protocols provide insight into teacher beliefs with respect to reform teaching methods and the extent to which they are able to use such methods in the classroom, interviews can provide a more in-depth understanding of the various issues associated with the execution of reform teaching. Being less structured than surveys or observation protocols, interviews are thus more flexible and allow the respondent to bring up new issues and guide the interview toward the most relevant ones (Axinn and Pearce 2006). Five teacher interview protocols were examined for this review. Two of these, the ITC-TIP and the OTIP have been mentioned in the previous section as they are associated with their own observation protocols (the ITC-COP and the OTOP, respectively). These protocols are intended to be used before or after the lesson being observed to obtain the instructor's perspective. While the OTIP has four items and is very open-ended, the ITC-TIP has more items, includes probes for some items, and is more specific. The *Views of Nature of Science Form B* (VNOS-B) interview is part of a protocol that includes a questionnaire and an apprenticeship and is intended to assess participants' knowledge of the process of scientific inquiry. The *Beliefs Questions* protocol and the *Teacher Beliefs Interview* (TBI) are both designed to obtain an understanding of teacher beliefs about their own teaching, as well as their students' learning. The TBI allows the development of a beliefs profile for each individual by classifying each response into one of five categories, while the Beliefs Question protocol has seven questions, each of which represents a dimension of beliefs. The fifth interview protocol, *the Levels of Use of an Innovation*, is part of the *Concerns-Based Adoption Model*, which has been widely used for over 25 years. The *Levels of Use* interview can be used with any innovation including reform teaching, and classifies a respondent into one of eight levels (three non-use and five use) based on their behaviour in response to the innovation under study and the extent to which they have mastered the innovation.

Discussion

As is evident from the results of this review, there is no simple answer as to which instrument or combination of instruments is 'best' for use in an evaluation of a professional development programme. The answer depends on the research question, purpose of the investigation, outcomes under study, methodological approach, stakeholders involved, cost of administration, and availability of resources. An instrument that may be appropriate for one study may not be suitable for another. Sometimes, it may even be advisable to use more than one survey or interview for an investigation.

In our investigation, we were interested in studying and evaluating the process of change in teacher beliefs and practice that happens after professional development has been administered, and also any concerns and issues teachers experience while trying to implement reform teaching in the classroom. After considering several instruments, we decided, for future projects with similar purposes, to use the TSI as our survey instrument, the RTOP as our observation protocol, and the TBI as our interview protocol. In this section, we discuss the factors that helped our decision.

Choice of the survey instrument

It is our conviction that following an effective professional development programme, described previously, teacher beliefs and confidence in their ability to execute reform teaching would undergo a change. Also, prior research has shown the relationship among teacher self-efficacy, teacher performance, and student performance (Ashton 1984). Increases in self-efficacy can be expected to have an impact on teachers' risk taking, which in turn affects their classroom choices usually displayed as reformed teaching practices (Ashton 1985). There have been an array of studies showing that as self-efficacy increases, so does the level of teacher performance, and that performance can be predicted from self-efficacy (Bandura 1977). Therefore, while measuring teacher beliefs, we decided to focus specifically on self-efficacy. This decision narrowed our choice of instruments to two, including the STEBI-B and the TSI.

The STEBI-B is a survey designed to measure self-efficacy, in this case, teachers' capabilities to produce designated levels of performance with regard to teaching. Based on the widely used Bandura's self-efficacy theory, Riggs and Enochs (1990) developed the STEBI-B which was subsequently modified by Bleicher (2004) for internal reliability and validity reasons. It has 23 items that make up two subscales – the *Science Teaching Outcome Expectancy* (STOE) scale and the *Personal Science Teaching Efficacy* (PSTE) scale. The STEBI-B is a well-established instrument and has been widely used for over a decade. Reliability and validity were established for the original instrument and reexamined by Bleicher in 2004.

The TSI is a relatively recent instrument (Smolleck, Zembal-Saul, and Yoder 2006), which measures teacher self-efficacy in reform settings. Like the STEBI-B, it has its theoretical basis in Bandura's theory of self-efficacy, and incorporates the five features that define teaching and learning science as inquiry across all grade levels, according to the National Science Education Standards (Smolleck, Zembal-Saul, and Yoder 2006) – namely that the learner: (1) engages in scientifically orientated questions, (2) gives priority to evidence in responding to questions, (3) formulates explanations from evidence, (4) connects explanations to scientific knowledge, and (5) communicates and justifies explanations. The TSI has 69 items and has undergone extensive instrument development. The TSI has established validity and reliability (Smolleck, Zembal-Saul, and Yoder 2006).

After comparing the two, we decided to adopt the TSI. A driving force for our choice of the TSI over the STEBI-B was that the TSI was specifically developed to study inquiry-based teaching whereas the STEBI-B was not. The STEBI-B was developed before the reform movement gained prominence, and some of the items do not reflect the ideals of reform teaching.

Choice of observation protocol

Two prominent criteria for our choice of observation protocol were: (1) the instrument had to have established psychometric properties and have been well-researched and widely used so that there was a body of literature available for us to study, and (2) the instrument needed to measure reform teaching from a developmental perspective. The ITC-COP and the RTOP met these criteria. Both are well regarded instruments and are designed to measure the extent of reform teaching in a classroom which is typically anticipated after participation in standards-based, science-focused, professional development.

The ITC-COP is an observation instrument developed by Horizon Research, Inc. for the observation of standards-based math and science instruction in K-12 classrooms. The instrument is 22 pages long with 67 items divided into three parts: (1) descriptive information, (2) influences in planning the lesson, and (3) summary and overall rating. Four domains of reformed teaching are evaluated – design, implementation, content, and classroom culture. Each of these parts has items that require ratings on a Likert-type scale, as well as narrative descriptions.

The RTOP is an observation instrument developed by the Arizona Collaborative for the Excellence in Preparation of Teachers (ACEPT) and is designed to measure the extent to which reformed teaching is evident in the classroom. The instrument has 25 items and is divided into five sections: (1) lesson design and implementation, (2) content: propositional knowledge, (3) content: procedural knowledge, (4) classroom culture: communicative interactions, and (5) classroom culture: student–teacher relationships.

The central factors that led to our selection of the RTOP were that its compact nature and online materials provided a cost-effective means to implementing the instrument for significant populations. As we discovered, and Henry, Murray, and Phillips (2007) describe, the RTOP can be used by most educators who are trained in its usage. A cursory review of the theoretical principles and item language indicates many commonalities, leading us to believe that it is reasonable (though not yet proven) that self-efficacy data collected through application of the TSI would correlate with RTOP scores.

Choice of interview protocol

Our goal in using a teacher interview protocol was to gain an in-depth understanding of the beliefs held by teachers as they used inquiry-based teaching in their classrooms and to also gain insight into any issues they face or concerns they have while executing reform teaching. Our purpose thus precluded the use of the ITC-TIP and the OTIP, which are closely related to their respective observation protocols. The LoU was not appropriate for our purposes as it measures the extent of mastery and use of reformed teaching rather than beliefs. Similarly, the VNOS-B was designed to measure participants' understanding of the nature of science and scientific inquiry and not beliefs.

We ultimately selected the TBI (Luft and Roehrig 2007) for our evaluation. The TBI is a seven question semi-structured interview, allowing the interviewer to ascertain the beliefs of the teachers and then use probes to further discuss them. The TBI includes the creation of a profile for each interviewee by classifying each response into one of five categories ranging from traditional to reform. Creating a profile for each teacher will not only help us examine trends among teachers but also make it possible to align that teacher's profile with results from the survey and observation protocols. Validity and reliability of the instrument have been examined and deemed acceptable.

Conclusion

This article was motivated by the absence of adequate literature available to assist with the location and selection of instruments that meet evaluation needs for K-12 science professional development efforts. As the emphasis on reform teaching and the associated professional development has grown, so has the need for quality evaluation of such programmes. The choice of instruments is a crucial part of any programme evaluation, and while the implementation of a multi-method approach aids with data triangulation and validity, the application of multiple instruments of differing methods leads to instrument alignment questions. However, if alignment can be verified, then the final evaluation will conclude in a comprehensive manner.

Of the range of survey, observation, and interview instruments available, most of them are self-contained instruments that have not been specifically developed to align with other instruments. The instruments that we selected, the TSI, RTOP, and TBI, have all been developed independently of each other; therefore, it is necessary to verify that they can be used in conjunction with each other. Our initial efforts have studied the theoretical backgrounds and face validity for the three selected instruments and early indications support the application of the three instruments in conjunction. Our next step is to apply the instruments to a population using the data to conduct a correlational study among the three instruments. Additional testing of alignment using more rigorous statistical methods is an issue that merits further research.

This article has established that the instruments available for use in the evaluation of K-12 science professional development vary in scope, outcomes measured, and methods and extent of psychometric properties – among other variables. We have compared and contrasted these factors for a selected sample of instruments available, and we have discussed the decision-making process that guided us in making our choice of instruments for a specific programme. It is our hope that this review provides a synopsis that may help evaluators and other educators in selecting instruments that suit their purposes.

References

Abd-El-Khalick, F. 1998. The influence of history of science courses on students' conceptions of nature of science. Unpublished doctoral dissertation. Oregon State University, Corvallis.

Abd-El-Khalick, F., R.L. Bell, and N.G. Lederman. 1998. The nature of science and instructional practice: Making the unnatural natural. *Science Education* 82, no. 4: 417–37.

Ashton, P. 1984. Teacher efficacy: A motivational paradigm for effective teacher education. *Journal of Teacher Education* 35, no. 5: 28–32.

Ashton, P. 1985. Motivation and teacher's sense of efficacy. In *Research on motivation in education: The classroom milieu,* vol. 2, ed. C. Ames and R. Ames, 141–74. Orlando, FL: Academic Press.

Axinn, W., and L. Pearce. 2006. *Mixed method data collection strategies.* New York: Cambridge University Press.

Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84, no. 2: 191–15.

Bandura, A. 1997. *Self-efficacy: The exercise of control.* New York: W.H. Freeman.

Beerer, K., and A.M. Bodzin. 2004. Promoting inquiry-based science instruction: The validation of the science teacher inquiry rubric (STIR). Paper presented at the Association for the Education of Teachers of Science (AETS) Annual Meeting, January 8–11, Nashville, TN.

Bell, R.L., L.M. Blair, B.A. Crawford, and N.G. Lederman. 2003. Just do it? The impact of a science apprenticeship program on high school students' understandings of the nature of science and scientific inquiry. *Journal of Research in Science Teaching* 40, no. 5: 487–09.

Bell, R.L., and N.G. Lederman. 2003. Understandings of the nature of science and decision making on science and technology based issues. *Science Education* 87, no. 3: 352–77.

Bell, R.L., N.G. Lederman, and F. Abd-El-Khalick. 2000. Developing and acting upon one's conception of the nature of science: A follow-up study. *Journal of Research in Science Teaching* 37: 563–81.

Birman, B.F., L. Desimone, M.S. Garet, and A.C. Porter. 2000. Designing professional development that works. *Educational Leadership* 57, no. 8: 28–33.

Bleicher, R.E. 2004. Revisiting the STEBI-B: Measuring self-efficacy in preservice elementary teachers. *School of Science and Mathematics* 104, no. 8: 383–91.

Bodzin, A.M., and W.M. Cates. 2002. Web-based inquiry for learning science (WBI) instrument manual. http://www.lehigh.edu/~amb4/wbi/wbi-v1_0.pdf (accessed August 15, 2008).

Cobern, W.W. 2000. *The Thinking about Science Survey Instrument (TSSI) – SLCSP 151.* Kalamazoo, MI: Scientific Literacy and Cultural Studies Project. http://www.wmich.edu/slcsp/slcsp151/tssi-v2.pdf (accessed September 18, 2008).

Crippen, K.J. 2008. The translation of motivation for teaching science to classroom practice in a large-scale professional development project. Paper presented at the Association of Science Teacher Education (ASTE) National Conference, January 12, in St. Louis, Missouri, USA.

Darling-Hammond, L. 1997. *Doing what matters most: Investing in quality teaching.* Kutztown, NY: National Commission on Teaching and America's Future.

Darling-Hammond, L., and G. Sykes, eds. 1999. *Teaching as a learning profession: Handbook of policy and practice.* San Francisco, CA: Jossey-Bass.

Dirr, P.J. n.d. Classroom observation protocols: Potential tools for measuring the impact of technology in the classroom. *Appalachian Technology in Education Consortium, Policy and Planning Series #104.* http://www.k12.wy.us/TCD/eval/classroom_protocols.pdf (accessed August 11, 2008).

Dole, J.A., and G.M. Sinatra. 1998. Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist* 33, nos. 2/3: 109–28.

Duttweiler, P.C., and U. McEvoy. 1999. Standards-based reform: What is missing? *Journal of At-Risk Issues* 5, no. 2: 3–10.

Feldman, A. 2000. Decision making in the practical domain: A model of practical conceptual change. *Science Education* 84, no. 5: 606–23.

Ford, M. 1992. *Motivating humans: Goals, emotions, and personal agency beliefs.* Newbury Park, CA: Sage.

Garet, M.S., A.C. Porter, L. Desimone, B.F. Birman, and K.S. Yoon. 2001. What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal* 38, no. 4: 915–45.

Glusac, T. 2008. Evaluating professional development. http://www.britishcouncil.org/serbia-feature-articles-glusac.doc (accessed August 15, 2008).

Gregoire, M. 2003. Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review* 15, no. 2: 147–79.

Guskey, T. 2000. *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

Hall, G.E., and S.M. Hord. 1987. *Change in schools: Facilitating the process: SUNY series in educational leadership.* Albany, NY: SUNY Press.

Henry, M.A., K.S. Murray, and K.A. Phillips. 2007. Meeting the challenge of STEM classroom observation in evaluating teacher development projects: A comparison of two widely used instruments. Paper presented at the American Evaluation Association Annual Conference, November 7–10, Baltimore, MD.

Kennedy, M. 1998. *Form and substance in inservice teacher education* (Research monograph no. 13). Madison, WI: University of Wisconsin, National Institute for Science Education Research.

Khourey-Bowers, C., and D.G. Simonis. 2004. Longitudinal study of middle grades chemistry professional development: Enhancement of personal science teaching self-efficacy and outcome expectancy. *Journal of Science Teacher Education* 15, no. 3: 175–95.

Lawrenz, F., D. Huffman, and K. Appeldoorn. 2002. *Classroom observation handbook.* Minneapolis, MN: University of Minnesota, College of Education and Human Development.

Lederman, N.G., and M. O'Malley. 1990. Students' perceptions of tentativeness in science: Development, use and sources of change. *Science Education* 74: 225–39.

Lederman, N.G., R.S. Schwartz, F. Abd-El-Khalick, and R.L. Bell. 2001. Pre-service teachers' understanding and teaching of the nature of science: An intervention study. *Canadian Journal of Science, Mathematics and Technology Education* 1: 135–60.

Lincoln, Y.S., and E.G. Guba. 1985. *Naturalistic inquiry.* Newbury Park, CA: Sage.

Luft, J.A. 2005. Beliefs questions. http://ret.fsu.edu/Files/Tools/Beliefs_Interview_Questions.htm (accessed August 15, 2008).

Luft, J.A., and G.H. Roehrig. 2007. Capturing science teachers' epistemological beliefs: The development of the teacher beliefs interview. *Electronic Journal of Science Education* 11, no. 2: 38–63. http://ejse.southwestern.edu (accessed April 12, 2007).

Lumpe, A.T., J.J. Haney, and C.M. Czerniak. 2000. Assessing teachers' beliefs about their science teaching context. *Journal of Research in Science Teaching* 37, no. 3: 275–92.

Moore, R.W., and R.L.H. Foy. 1997. The scientific attitude inventory: A revision (SAI II). *Journal of Research in Science Teaching* 34, no. 4: 327–36.

National Education Goals Panel. 1999. *The National Education Goals report: Building a nation of learners.* Washington, DC: US Government Printing Office.

National Research Council. 2000. *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy of Sciences.

Patton, M.Q. 2002. *Qualitative research and evaluation methods.* Newbury Park, CA: Sage.

Piburn, M., and D. Sawada. 2000. *Reformed teaching observation protocol (RTOP) reference manual: Technical report* (Report no. ACEPT-TR-IN00-32). Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers (ERIC Document Reproduction Service No. ED447205).

Posnanski, T.J. 2002. Professional development programs for elementary science teachers: An analysis of teacher self-efficacy beliefs and a professional development model. *Journal of Science Teacher Education* 13, 189–30.

Ravitz, J.L., H.J. Becker, and Y. Wong. 2000. *Constructivist-compatible beliefs and practices among U.S. teachers: Teaching, learning and computing: 1998 Survey report no. 4* (ED 445 657). Irvine, CA: Center for Research on Information Technology and Organizations, University of California-Irvine.

Resnick, L.B. 2005. Teaching teachers: Professional development to improve student achievement. *Research Points* 3, no. 1: 1–4.

Riggs, I.M., and L.G. Enochs. 1990. Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics* 90, no. 8: 694–06.

Sampson, V., and A. Benton. 2006. Development and validation of the beliefs about reformed science teaching and learning (BARSTL) questionnaire. http://theaste.org/publications/proceedings/2006proceedings/Sampson_ASTE2006%201%20.htm (accessed August 20, 2008).

Smolleck, L.D., C. Zembal-Saul, and E.P. Yoder. 2006. The development and validation of an instrument to measure preservice teachers' self-efficacy in regard to the teaching of science as inquiry. *Journal of Science Teacher Education* 17, no. 2: 137–63.

Southerland, S., S. Sowell, M. Kahveci, D.E. Granger, and O. Gaede. 2006. *Working to measure the impact of professional development activities: Developing an instrument to quantify pedagogical discontentment.* San Francisco, CA: Florida State University, Research Experience for Teachers. http://ret.fsu.edu/Files/Tools/Discontentment_Instrument.doc (accessed August 6, 2008).

Strike, K.A., and G.J. Posner. 1992. A revisionist theory of conceptual change. In *Philosophy of science, cognitive psychology and educational theory and practice,* ed. R. Duschl and R. Hamilton, 147–76. Albany, NY: SUNY Press.

Taylor, P.C., B.J. Fraser, and D.L. Fisher. 1997. Monitoring constructivist learning environments. *International Journal of Educational Research* 27, no. 4: 293–02.

US Department of Education. 2008. Mathematics and science partnerships. http://www.ed.gov/programs/mathsci/index.html (accessed January 20, 2009).

Von Glasersfeld, E. 1989. Cognition, construction of knowledge and teaching. *Synthese* 80, no. 1: 121–40.

Wainwright, C.L., L.B. Flick, and P. Morrell. 2003. Development of instruments for assessment of instructional practices in standards-based teaching. *Journal of Mathematics and Science: Collaborative Explorations* 6, no. 1: 21–46.

Wee, B., D. Shepardson, J. Fast, and J. Harbor. 2007. Teaching and learning about inquiry: Insights and challenges in professional development. *Journal of Science Teacher Education* 18: 63–89.

Weiss, I.R., J.D. Pasley, P.S. Smith, E.R. Banilower, and D.J. Heck. 2003. *Looking inside the classroom: A study of k-12 science and mathematics in the United States.* Chapel Hill, NC: Horizon Research.