Review of Education
An International Journal of Major Studies in Education
BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION
Check for updates

Review of Education
Vol. 9, No. 1, February 2021, pp. 81–120

DOI: 10.1002/rev3.3235

A systematic review of observation protocols used in postsecondary STEM classrooms

Saira Anwar[1] and Muhsin Menekse[2]
[1]*Department of Engineering Education, University of Florida, Gainesville, FL, USA,*
[2]*School of Engineering Education, Department of Curriculum and Instruction, Purdue University, West Lafayette, IN, USA*

Prior research studies have extensively used different classroom observation protocols to identify the characteristics of the lecture and instructional methods used by course instructors, to observe student and instructor behaviours, to evaluate the fidelity of certain implementations, and to understand classroom dynamics. This systematic review provides a synthesis and comparison of the features and dimensions of commonly used observation protocols in postsecondary (undergraduate or college) classrooms. This study identifies eight observation protocols, which are: Reformed Teaching Observation Protocol, Oregon Teacher Observation Protocol, VaNTH Observation System, Cooperative Learning Observation Protocol, Teaching Dimension Observation Protocol, Classroom Observation Protocol for the Undergraduate STEM, Classroom Interactive Engagement Observation Protocol, and Student Resistance and Instructional Practices. Based on 35 articles included in the full review, we present an overview description of each protocol. Further, we describe the strength and limitations of these protocols using deductive thematic and content analysis. Also, we evaluate these protocols by using the assessment triangle approach. The study advances the existing literature by providing usability details and suggestions to use each protocol. This review study also enhances the literature in science, technology, engineering and mathematics (STEM) education as the first study that evaluates all these observation protocols and provides a reference point for future researchers.

Keywords observation methods, observation protocols, postsecondary classrooms, STEM education, systematic review.

Introduction

Education studies have extensively explored the effectiveness of evidence-based instructional practices and reforms on science, technology, engineering, and mathematics (STEM) courses (Borrego & Henderson, 2014). The increased focus resulted from national calls for improvement in STEM education, instructional practices, and students' quality (e.g., ASEE, 2012; PCAST, 2012). To ensure the implementation and demonstration of evidence-based practices, researchers identified various methods to evaluate students' behaviours and instructional practices (AAAS, 2013). Commonly used methods include conducting course evaluations or performance assessments, surveys and rubrics, observing classes, writing detailed narratives of classes (field notes), videotaping or recording lectures, peer observations for

Corresponding author. Saira Anwar, Department of Engineering Education, University of Florida, Nuclear Science Building, Gainesville, IN 32611, USA. Email: sairaanwar@ufl.edu

feedback, faculty/student interviews, and student outcomes (Berk, 2005). However, researchers advocate using observations over other methods as a more comprehensive method for classroom practices, teaching evaluations, and empirical research (e.g., Pianta & Hamre, 2009).

Broadly speaking, in classroom observations, researchers make notes of instructors and student behaviours in real-time, or from a videotaped/recorded lesson (Hora, 2015). However, it is vital for effective and reliable observation to use an appropriate tool or observation protocol which provides reliable data (Hora, 2015). Also, the observation protocol should have the capability to capture the research's theoretical perspective (Lund *et al.*, 2015). For example, the protocol may capture details of effective teaching, classroom dynamic, student behaviours, instructional methods, and student-teacher interactions (Ebert-May *et al.*, 2011).

Studies have used observation protocols in classroom settings across postsecondary (includes universities, colleges, and trade or vocational schools) and K-12 (kindergarten to 12th grade) grade levels (e.g., Sawada *et al.*, 2002; Tolentino *et al.*, 2009; Campbell *et al.*, 2010; Whiteside *et al.*, 2010; Amrein-Beardsley & Popp, 2012). These observation protocols can be broadly classified into two broad categories: (1) open-ended observation protocols, which are unstructured accounts of observer's detail notes), and (2) structured observation protocols, which are designed with a set of predetermined and defined statements (Pretzlik, 1994).

While prior research studies emphasised the importance of observation protocols, less information is available to decide which protocol is most suitable for any given context. This selection of appropriate protocol is particularly difficult due to each protocol's embedded goals and unique features. For example, some protocols focus on documenting the students' engagement and resistance; others may only capture the instructors' teaching strategies. Also, as these protocols are contextually designed (for a particular setting, grade-level, and with some STEM disciplines), it is important to have a complete understanding (Wainwright *et al.*, 2003). Studies have not focused on categorising these protocols based on different research perspectives, observation categories, contextual differences, and validity of protocols (Wainwright *et al.*, 2003). The lack of understanding and knowledge of specific goals and unique features of protocols make the selection harder for researchers. A limited research has focused on comparing the observation protocols and their features. Also, scarcity of literature evaluates the validity of available protocols, which is an important aspect of researchers' decision to use the protocol.

Considering the need to evaluate and compare different observation protocols based on multiple factors, we focused on developing an understanding of the most commonly used structured protocols in this systematic review. We focused on the features of commonly used observation protocols and synthesised the literature on common ways the protocol was used. This literature review is conducted with two goals: (1) to provide a unified point of contact to future researchers, which aids in the understanding of a comparative view of different observation protocols; (2) to present empirical evidence about the validity of the protocols. We believe that this literature review will help the researchers to decide which protocol to use, based on their needs. More specifically, the following questions guided the study:

RQ1) What are the most common classroom observation protocols designed for postsecondary classrooms?
RQ2) What are the primary purposes, strengths, and limitations of each protocol used in postsecondary classrooms?
RQ3) What principles were used to validate each observation protocol?

This review paper on observation protocols covers those protocols published during 1990–2020 and used in a postsecondary undergraduate STEM context. Moreover, as the student-faculty ratios have significantly grown over the past decades in most STEM courses, resulting in large class sizes, we focused on the observation protocols that have been used in large lectures. Considering the need for the research goals and questions, we used a systematic review approach and classified 35 studies. In each paper, we analysed the context, validity, and usage details of an observation protocol. Further, we synthesised these observation protocols on their characteristics, strengths and limitations. This review also evaluates protocols based on their validity and reliability using a standardised method of assessment triangle.

Literature review

Prior research studies have indicated the focus of policymakers and national reforms towards understanding the link between classroom practices and students' outcomes (Pianta & Hamre, 2009). For a reliable understanding of these links, researchers have focused on research tools, including classroom observations (Hora, 2015). The classroom observations helped researchers document instructional practices, and students' experiences towards various teaching methods (Hora, 2015). Over the years, one common mechanism of conducting classroom observations is through the use of observation protocols (Pretzlik, 1994). Initially, these techniques were more rigorously used in K-12 settings (e.g., Pianta & Hamre, 2009). However, increased emphasis on postsecondary instruction (e.g., PCAST, 2012) resulted in the design and use of observation protocols in postsecondary (undergraduate, college or vocational) STEM courses.

Education studies have extensively used observation protocols to explore students' in-class behaviours, instructors' teaching strategies, classroom dynamics, and assure the fidelity of implementations (Ross *et al.*, 2004; Huntley, 2009). Prior research studies have described various classifications of the observation protocols. For example, these protocols can be divided into types according to the nature of their recording. Also, these observation protocols can be categorised based on the nature of the questions in the protocol.

Based on the type of recording, the observation protocols can be divided into two types: (1) holistic protocols, which require the coding of each item for the entire class period, and (2) segmented protocols, which require evaluating each item in time segments (AAAS, 2013). Further, the observation protocols can generally be classified into two broad categories: (1) open-ended observation protocols (unstructured), and (2) structured observation protocols (Pretzlik, 1994). The usage of these two types was mostly dependent on the research questions, the paradigm of the study (Mulhall, 2003; Smith *et al.*, 2013), and the researchers' philosophy (Mulhall, 2003).

The open-ended or unstructured protocols were developed and used based on the importance of context, situation, and co-construction of knowledge in research (Mulhall, 2003). Such protocols provide autonomy to the observer. Consequently, these protocols were often used by the proponents of interpretivism and naturalistic paradigms (Mulhall, 2003). The observer can provide comments on the situation, environment or behaviours of the participants (Millis, 1992; Parahoo, 2014). The observer role is defined as attending the setting, taking notes, and responding to open-ended questions (Smith *et al.*, 2013). These methods are best used to analyse a particular situation, provide feedback, or study the detailed accounts of participants' behaviour, attitudes and experiences (Mulhall, 2003).

Alternatively, researchers often used structured protocols following the positivistic approach (Mulhall, 2003). These protocols are standardised tools and designed with a set of statements or a set of predetermined codes. These tools are based on specific taxonomies derived from a known theory (Mulhall, 2003). In general, structured protocols are designed to evaluate the occurrence, frequency and intensity of a particular behaviour or characteristic (Parahoo, 2014). The observer role is to make judgements by giving a value (e.g., Likert scale-based values) or to assign a specific code to the occurrence (Smith *et al.*, 2013; Parahoo, 2014). These observation protocols have been used and developed to study various parameters of educational environments which may help to understand instructional effectiveness. For example, observation protocols have been used to understand the course of events in formal and informal learning settings, to explore the mechanism of interactions that happen between students – student, student-teacher, and teacher – course material, or to investigate participants' engagement (Brophy & Good, 1986; Stallings & Mohlman, 1988; Waxman & Padron, 2004). Further, these protocols can be used to evaluate the instructors' teaching pedagogies, methods and instructional strategies.

Although much research is conducted in the design and implementation of various observation protocols, the less is known to the user that which protocol is more appropriate in any given context. In most cases, the research team for developing each new protocol described the need for a new protocol in lieu of the absence of features in the existing protocol. The collective comparison research is sparse on the subject. Partly in response to the lack of comparative research on the existing observation protocol, this systematic review provides a rigorous description of each protocol from various angles of strengths, limitations, usage suggestions and validity.

Research methods

This study used the systematic literature review methodology to search, review, analyse, and synthesise the literature. We followed Borrego *et al.*'s (2015) guidelines to search the databases, select the keywords and studies, code the studies, and synthesise the articles.

Search databases

We focused on searching the three databases: (1) Proquest Research Library, (2) IEEE Xplore and (3) ERIC. The search was performed in the summer of 2020. Once

Table 1. The search protocol for the review

Database	Search Protocol
ProQuest Research Library	Search String: 'Observation Protocol' AND (Postsecondary OR Undergraduate) AND (STEM OR Science OR Technology OR Engineering OR Mathematics) AND Large AND (Class OR Course) Searched in: Full text and peer-reviewed articles
IEEE Xplore	Search String: 'Observation Protocol' Searched in: Full text only We narrowed down the papers based on other criteria during the screening process
ERIC	Search String: 'Observation Protocol' AND (Postsecondary OR Undergraduate) AND (STEM OR Science OR Technology OR Engineering OR Mathematics) AND Large AND (Class OR Course) Searched in: Peer-reviewed only
Other sources	We included the seminal papers of each protocol written by the authors of the protocol.

we determined which structured observation protocol was used, we looked for the seminal papers and included them in the pool of studies. The search protocol is described in Table 1.

All sources resulted in a total of 124 papers. Each database resulted in 80, 19, 6 and 19 studies, respectively.

Selection criteria and process

In addition to the search protocol and removing duplicates, we used inclusion and exclusion criteria to screen the studies. We included studies that focused on structured protocols. Specifically, In this literature review, we used studies that either described the creation (introduce) process of the new structured observation protocol or have used a structured observation protocol for classroom observation. We also included papers which focused on describing the results of the classroom observation and we excluded articles that were written in any language other than English. We further excluded articles which did not focus on an undergraduate STEM course setting or were not written as a full paper (e.g., editorial, or work in progress papers).

For ensuring a high-quality review, Figure 1 shows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (Moher, Liberati, Tetzlaff, Altman, & Prisma Group, 2009).

Based on the above criteria, we excluded 89 papers due to their nonrelevance to the current literature review, and analysed 35 papers. Please see Appendix 1 for the complete list of the papers.

Data analysis

First, we categorised each study according to the primary research goal of the article and according to two distinct categories: (1) it introduces the observation protocol or

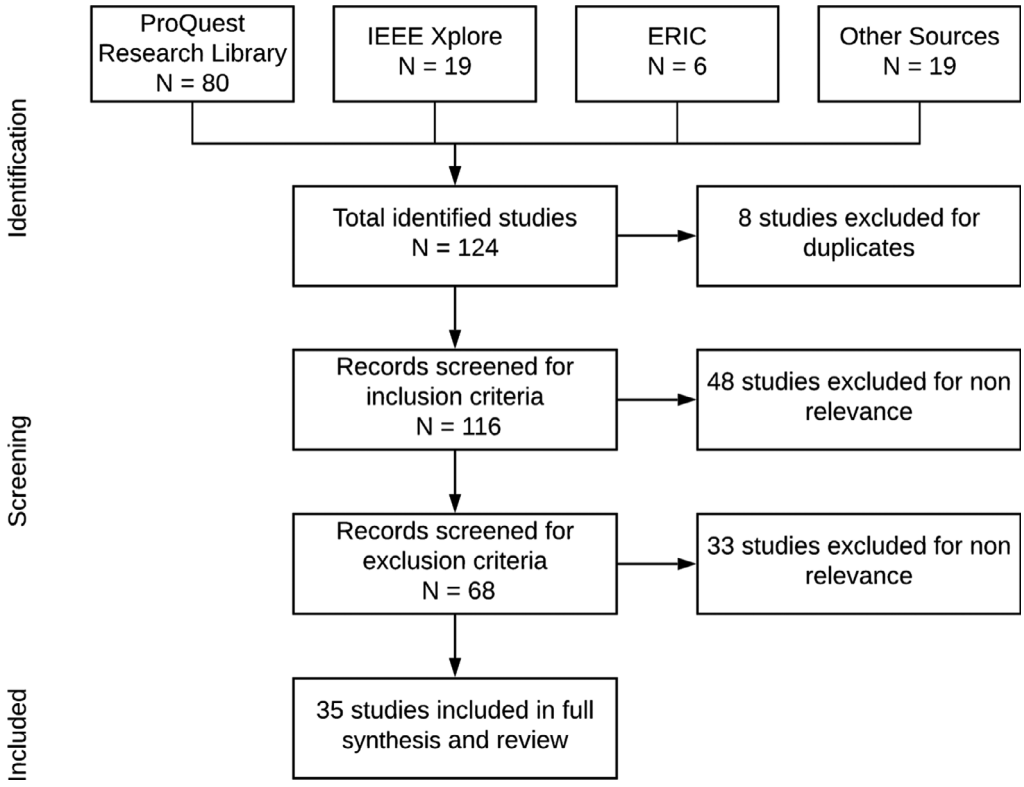

Figure 1. PRISMA – Flow of information through stages (Moher, Liberati, Tetzlaff, & Altman, 2009)

is one of the seminal papers of the protocol, and (2) it uses an existing structured observation protocol for classroom observation. Categorising in this way helped to identify the most commonly used observation protocols in postsecondary STEM classes. We then focused on understanding and synthesising literature to overview the protocols – their strengths and limitations – by using content analysis and deductive thematic analysis. Also, we highlighted the suggestions for using each protocol. Further, we used the assessment triangle to understand the validity measures of each protocol.

Results

We classify the results in three sections: (1) common observation protocols; (2) overview, strengths, limitations and suggestions to use for each protocol; and (3) validity measures in each protocol.

Common observation protocols

We reviewed the literature according to the goal of each included research study. We categorised each paper into either an introduction to the observation protocol, or one

Table 2. The frequencies distribution for two categories

Observation protocol	Frequency of studies	
	Introduces	Uses
RTOP	2	7
OTOP	1	1
VOS	1	1
CLOP	1	0*
TDOP	2	6
COPUS	1	7
CIEOP	1	1
StRIP	2	1
Total	11	24

*CLOP is used in Rivera, N. (2013). *Cooperative Learning in a Community College Setting: Developmental Coursework in Mathematics* (doctoral dissertation, Arizona State University).

which used an existing protocol. Also, we reviewed the literature to identify the commonly used observation protocols in postsecondary STEM classrooms. We found a total of eight protocols which have been used in STEM classrooms. These protocols are a Reformed Teaching Observation Protocol (RTOP), Oregon Teacher Observation Protocol (OTOP), VaNTH Observation System (VOS), Cooperative Learning Observation Protocol (CLOP), Teaching Dimension Observation Protocol (TDOP), Classroom Observation Protocol for the Undergraduate STEM (COPUS), Classroom Interactive Engagement Observation Protocol (CIEOP), and Student Resistance and Instructional Practices (StRIP). Table 2 shows the frequency distribution of the common protocols.

Overview of the protocols

Reformed Teaching Observation Protocol (RTOP). RTOP was developed by the Evaluation Facilitation Group (EFG) at Arizona State University to measure 'reformed' teaching and practices (Piburn & Sawada, 2000; Sawada *et al.*, 2002). The earlier implementation of the observation protocol was on mathematics and science classrooms in K-12 settings. Later, the protocol was also used in undergraduate classes (e.g., Hilpert & Husman, 2017; Frost *et al.*, 2018). The tool was designed in a quantitative Likert-scale style to evaluate faculty effectiveness and their teaching on reformed principles. These reformed principles include equity, curriculum, teaching, learning, assessment and technology. The protocol also contains the standards of problem-solving, reasoning and proof, communication, connections and representation. The protocol requires the observer to record field notes and provide personalised observation in order to document each class (Wainwright *et al.*, 2003).

Along with the field notes, the RTOP requires evaluation of the faculty on 25 items. These items correspond to classroom practices, which are: (1) lesson design and implementation (five items), (2) content (10 items), and (3) classroom culture (10 items) (Piburn and Sawada, 2000; Sawada *et al.*, 2002). The content items have a further division of procedural and propositional knowledge. The classroom culture items

have also been subdivided into two additional categories as communicative interactions and student-teacher relationships. The observer records their observation on all 25 items. The recorded observation determines the extent of each practice in the classroom on the scale of a five-point Likert scale from 0 ('never occurred') to 4 ('very descriptive').

Strengths—The tool is extensively tested for the fidelity of the implementation of specific reforms in K12 mathematics and science courses (e.g., Lawson *et al.*, 2002; Banchi, 2009; Liang *et al.*, 2012), as well as in undergraduate STEM courses (e.g., Middleton *et al.*, 2015; Hilbert & Husman, 2017). Prior studies have widely used RTOP for various purposes of observation which include: instructor and student behaviours in classrooms (e.g., Adamson *et al.*, 2003; Rushton *et al.*, 2011; Teasdale *et al.*, 2017), and for faculty evaluations (e.g., Hilpert & Husman, 2017), to understand the effectiveness of a professional development intervention on instructors' teaching practices (e.g., Smith *et al.*, 2015; Teasdale *et al.*, 2014) or peer review purposes (e.g., Ebert-May *et al.*, 2011; Frost *et al.*, 2018).

Limitations—The RTOP can be best reported as an instructor-focused tool (Shekhar *et al.*, 2015), with predefined static questions and one-time determination of instructor behaviours. This determination of behaviours has led to the use of RTOP as a faculty evaluation tool to assess the quality of the teaching (Hora, 2015) and the effectiveness of professional development workshops (Adamson *et al.*, 2003; Ebert-May *et al.*, 2011, 2015; Frost *et al.*, 2018). Studies report that while placing much focus on instructor behaviours, RTOP fails to capture the students' learning and engagement data (Wainwright *et al.*, 2003; Shekhar *et al.*, 2015). The protocol focuses less on content taught in class (Wainwright *et al.*, 2003; Hora, 2015). The tool also requires extensive training for observers (Sawada *et al.*, 2000). It is difficult for raters to distinguish between values of the Likert scale, especially in the absence of any standardised rubric describing the rationale for a specific value. Further, RTOP judgement is very observational, which is suited for institutional data. On 25 practices, the sum of all the sub-scores produces an overall score of 0 to100 (Liang *et al.*, 2012). Generally, only the overall score is shared with the faculty, and no item-based information is provided. This limitation of not sharing item-based information makes it hard for the faculty to reflect and improve course-based practices.

Suggestions for using the RTOP—The RTOP can be used to evaluate the fidelity of an implementation of a particular reform, as well as to evaluate instructor behaviours, and assess the faculty's instructional strategies. In the absence of a rubric, it is mandatory to train the observers on the Likert scale values by using the protocol manual (Piburn & Sawada, 2000). The protocol users have also advised on a training need in light of the lack of a 'not applicable' option (Henry *et al.*, 2007). It is further essential that observers know the subject domain in order to rate some of the items appropriately. For example, items like 'The teacher had a solid grasp of the subject matter content inherent in the lesson' are hard to evaluate without knowing the particular concepts of the subject area. The authors of the protocol have recommended having more than one observer in a lecture for classroom usage (Piburn & Sawada, 2000).

Overall, we recommend that prospective users of RTOP should have a clear understanding of each item before using it in large lectures. It would be valuable to create a rubric to differentiate what absolute ratings on each item indicate. Also, establishing interrater reliability between multiple raters is a must to use RTOP reliably in large lectures.

Oregon Teacher Observation Protocol (OTOP). OTOP was developed under the NSF-funded grant Oregon Collaborative for Excellence in the Preparation of Teachers (OCEPT). Due to maintaining the grant reference, the protocol is also referred to as OCEPT Classroom Observation Protocol (Morrell *et al.*, 2004). OCEPT's goal was to foster innovations in the instruction and assessment of mathematics and science courses (Wainwright *et al.*, 2003). The OTOP protocol was designed as part of the programme's outcome research study to determine the impact of professional development and assess standard-based instructional practices. The instrument was designed in a quantitative 5-point Likert scale style (N/O, 1–4) and had ten possible indicators. These indicators are used to evaluate the impact of reformed principles-based professional development on instructional practices, and student behaviours. N/O indicated 'not observed' and numbers indicated the frequency of the occurrence. Each indicator has a descriptive statement, a focus and rubric style descriptive information for the instructor. Nine out of 10 items also had descriptive information for recording student behaviours. The indicators' focus includes habits of mind, metacognition, student discourse and collaboration, rigorously challenged ideas, student preconceptions and misconceptions, conceptual thinking, divergent thinking, interdisciplinary connections, pedagogical content knowledge, and multiple representations of concepts (Wainwright *et al.*, 2003). The protocol is to be filled in after the class, based on the narrative field notes recorded during the class (Wainwright *et al.*, 2003).

Strengths—OTOP has been used and tested for comprehensively documenting the complexities of constructivist teaching from both a learning and an instructional standpoint (Morrell *et al.*, 2004). Also, the protocol is used in both K-12 (Morrell *et al.*, 2004) and undergraduate contexts (Wainwright *et al.*, 2004). Studies have used OTOP as a descriptive tool for generating a holistic profile of what is happening across the instructional setting (e.g., Wainwright *et al.*, 2004). The protocol has kept observation categories to 10 practices, which are easy to follow and remember (Wainwright *et al.*, 2003). Also, the protocol provides a specific description with each item, which aids in the understanding of each item's intent in a clear manner. The protocol provides a detailed account of reflections on instructional practice and students' in-class behaviours with instructors and peers. These reflections can be treated as feedback to instructors for the improvement of instructional practices.

Limitations—Although OTOP focuses on both instructor practices and student behaviours, the set of predefined static questions records only a few aspects of student behaviours. For example, similar to RTOP, the protocol does not account for student resistance or disengagement to tasks. Similar to RTOP, the goal of OTOP is focused on the evaluation of the instructional setting and the effectiveness of professional

development only (e.g., Wainwright *et al.*, 2004). Although student behaviour is a useful dimension, the focus neglects aspects of classroom dynamics. The authors of the protocol reported 100% agreement on eight out of 10 practices, and 57% and 71% for the remaining two practices using videotaped lectures (Wainwright *et al.*, 2003). However, the information lacks hours of necessary training for such an agreement.

Suggestions for using the OTOP—The OTOP can be used as a descriptive tool to generate a profile for what is happening in the class, with both instructional practices and student behaviours. However, the protocol relies heavily on field notes. The instrument sheet is filled after the class, based on the field notes. This heavy reliance on field notes requires observer training on writing descriptive narratives for all ten practices and understanding each category's scale. Also, in the case of multiple observers, it is important to establish interrater reliability between observers. One training hitch could be an observer's nonunderstanding of indicators. For example, Wainwright *et al.* (2003) reported observers' lack of understanding about metacognition and misconception/preconception indicators. For an appropriate description of the instructional setting, it is recommended to categorically use OTOP numerical values while analysing the data (Wainwright *et al.*, 2004).

VaNTH Observation System (VOS). The VOS was developed by Harris and Cox (2003), and Cox and Cordray (2008) as a discipline-specific instrument for bioengineering. The VOS addressed the need for an assessment tool which meets the Accreditation Board for Engineering and Technology (ABET) educational standard for engineering students. VOS's primary purpose is to document the instructional strategies in bioengineering classrooms. and assess the presence and absence of How People Learn (HPL)-practices (Bransford *et al.*, 1999) in classrooms (Harris & Cox, 2003). The protocol uses four lenses of HPL, as knowledge-centred, learner-centred, assessment-centred, and community-centred (Bransford *et al.*, 1999). The protocol has four components to indicate the methods of data collection (Harris & Cox, 2003). First, classroom interaction observation (CIO), which is recorded at three-minute intervals in a string form of who–to whom–said/did what–how–with what media. In this string, the 'how' specifically addresses the points of centredness, as described in HPL. Second, student engagement observation (SEO), which periodically compares the ratio of engaged students with the total number of present students in the class, using the categories of definitely engaged or probably engaged. The protocol also determines the engagement medium, such as with professors, independently, or with the media. Third, within the narrative notes (NN), five types of narratives are used: professor lectures, professor questions, professor guides problem solving, student leads the class, and organisation. Fourth, global ratings (GR) are recorded after the observed class to indicate the use of HPL-based pedagogies and to describe teacher and student actions of cognitive indication, student understanding and lesson engagement.

Strengths—The protocol is designed to record the instructional pedagogies and to be a useful tool to provide feedback to instructors on their teaching strategies (Cox

et al., 2011). The use of HPL dimensions can help to assess the type of course as assessment-centred, knowledge-centred, learner-centred or community-centred. The VOS codes on student engagement also allow observers to record off-task student behaviours, such as social media usage or any other distraction. Furthermore, the codes document the type of questions asked by the students (Harris & Cox, 2003).

Limitations—The protocol was designed explicitly for bioengineering courses and is not used outside of bioengineering classes. Use in other STEM courses probably indicates different HPL indexes. The protocol requires extensive training for observers. Also, the broad nature of the codes requires videotaping the lecture (Harris & Cox, 2003). Harris & Cox (2003) recommended revising the CIO and HPL index because of the coding scheme's intricacy and complexity. Also, in the protocol, every activity that is not part of classroom organisation is coded as 'knowledge-centred', which is different from the actual definition of knowledge centredness of helping students to understand (Bransford *et al.*, 1999). The protocol has a more focused approach towards recording CIO than the remaining three aspects. For example, student engagement is based on the observer's ability to count the heads accurately in real-time, not an easy task in large classrooms. Shekhar *et al.* (2015) argued that the protocol's engagement definition includes students' note-taking behaviours and listening to the lecture, which may not necessarily represent the same thing.

Suggestions for using the protocol—The VOS protocol is used for studying instructional strategies, students' behaviours, and engagement in the bioengineering classroom. The protocol recording time and codes needs some tailoring based on the structure and logistics of the class and maybe for other disciplines. Also, the protocol requires extensive training on HPL aspects. In SEO, the same aspect can be recorded in multiple string formats, and it is hard to distinguish between them while recording, which also confirms the need for training to know the difference. For example, a professor who is talking to students by using slides or a professor who is using media for instruction may mean the same thing in some contexts and this needs clarity when recording. The authors of the protocol reported the use of operational definitions of HPL aspects (Harris & Cox, 2003), which is probably essential during training. To record SEO correctly, it is important to have more than one observer in a large classroom. In the original study, the authors indicated the four observers' presence in a class size of 60+ students (Harris and Cox, 2003).

Cooperative Learning Observation Protocol (CLOP). The CLOP was developed by using the cooperative learning theory (Slavin, 1984) and its five essential elements (Kern *et al.*, 2007). These elements are collectively called PIGS-Face. These elements are: (1) positive interdependence – all team members will work towards a common goal for the benefit of both individuals and the team (P); (2) individual accountability – individual members take responsibility and participate in the task (I); (3) group processing – use ways to improve the team processes and maximise learning (G); (4) social skills – engagement of appropriate interpersonal skills (S); and (5) promote each other's success through a supporting environment with face to face interaction (F) (Johnson, Johnson, and Smith, 1998; Johnson and Johnson, 1999). The

social skills element was removed from the later version of the protocol, based on its overlap with other elements. This protocol was based on the protocol developed under the NSF programme titled Collaboratives for Excellence in Teacher Preparation, called CETP (Lawrenz, Huffman, Appeldoorn, and Sun, 2002). The CLOP protocol has a specific focus on delineating students' cooperative learning skills while they are engaged in small group tasks. This criterion-referenced instrument evaluates students' interactions on four elements and records the students' engagement level on the scales of low, medium and high, and is not observed by using a specially designed rubric.

Strengths—The protocol can record instances of cooperative learning and can be used to rate participants' behaviours on a rating scale (Luo, 2017). Also, the protocol documents the elements of the level of cooperation, along with the level of students' engagement (Luo, 2017). The protocol is relatively short and allows the observer to take extensive field notes while evaluating cooperative interactions in classes. Also, CLOP is used as a useful tool when applied to a community of practice (Maresca *et al.*, 2014).

Limitations—There are some limitations to the CLOP. The protocol may not be suitable for use in large classrooms with multiple teams, as the CLOP was designed to observe team-based interactions. Also, the protocol may require multiple observers in the presence of more than one team to fully record the engagement and interactions in a group, or with team tasks (e.g., Rivera, 2013). The protocol also needs other sources of process data, such as videotaping the interactions, specifically in the case of multiple teams and their interactions (e.g., Rivera, 2013). Furthermore, the header of the protocol requires the class's demographic information and instructional context (Kern *et al.*, 2007), which require in-advance information from the instructor before visiting the classroom to observe.

Suggestions for using the protocol—The protocol is best used to evaluate students' group activities, engagement and interactions. It is best used in small classrooms or by focusing on one or two teams in a large class. More than one observer is needed to focus on each team in a large classroom context. Although the training requirement is not mentioned in the protocol, it is essential to train multiple observers, so they record similar aspects in field notes for each team. Users of the protocol suggested the understanding and use of the Cooperative Learning Observation Guide (CLOG) (Rivera, 2013). Further, students' engagement level in each element is subjective to the observer, and training is required to remove the bias. Literature supports the need for a post-observation meeting between observers to discuss their opinion of the activity and students' behaviours (Rivera, 2013).

Teaching Dimension Observation Protocol (TDOP). TDOP was developed by using the instructional systems-of-practice framework (Hora & Ferrare, 2013; Hora, 2015). The protocol has a specific focus on course planning and classroom instruction. TDOP was designed to evaluate classroom situations, as well as actors, and produced artefacts based on six aspects of classroom dynamics. These dimensions are further

categorised into the 'basic dimensions' and 'optional dimensions' of teaching. The basic dimensions include instructional practices, student-teacher dialogue and instructional technology. The optional dimensions help with understanding the additional details in classes. The optional dimensions include potential student cognitive engagement, pedagogical strategies and students' time on task. By using TDOP, each of these dimensions can be observed in two-minute intervals and can be recorded by assigning a set of associated codes. The primary purpose of the protocol is to determine the fidelity of a course, planning implementation in actual classroom environments with the following possible outcome measures: (1) instructional techniques used for teaching, (2) formative feedback for professional development, (3) study of the effect of instructional interventions, and (4) study of the variation in teaching practices across lectures or between groups ('Teaching Dimension Observation Protocol', 2010).

Strengths—TDOP has multiple strengths. First, the protocol is very detailed and provides various codes to capture each dimension with an equal focus on both students and instructors (Shekhar *et al.*, 2015). Second, with the two-minute recording interval, the protocol can record temporal fluctuations (Hora, 2015). Third, it is nonevaluative and allows the instructor to not feel evaluated on the quality (Hora & Ferrare, 2013). Fourth, it can be a valuable tool for providing formative feedback to teachers to improve their planning and implementation. This aspect of improving strategies is evident in studies that used the protocol and determined instructors' use of teaching methods (e.g., Finelli *et al.*, 2014; Hora *et al.*, 2017).

Furthermore, TDOP provides a rigorous method of measuring teaching as an empirical phenomenon (Hora & Ferrare, 2013). Finally, it can be used to observe the fidelity of implementation for a planned lecture. Although the protocol is not for evaluating the instruction's quality, the combination of the codes can provide the details on the desired discipline-specific practices (Hora & Ferrare, 2013).

Limitations—There are some limitations to TDOP as well. The high number of codes (i.e. 46 codes) requires intensive training and practice. It is hard for an observer to refer back to descriptions while observing. The protocol also fails to capture actual student engagement, attention, and what they are doing during activities (Moore, 2017; Smith *et al.*, 2013). For example, if a teacher asks students to have small group discussions, but students get engaged in some other work, TDOP codes do not allow observers to capture such instances. It is more focused on instructors' directions and behaviours, instead of focusing on students' actual behaviours (Shekhar *et al.*, 2015).

Suggestions for using the protocol—TDOP is well suited to record (1) what instructional practices have been used, (2) how a specific lecture was taught, and (3) common behaviours of the whole class (Smith *et al.*, 2013). The TDOP data takes the form of frequencies of each code, which can help understand the variation between students' passive and active roles. The protocol can also effectively shed light on an instructor's approach towards lectures (e.g., student-centred or content-centred), as well as characterise differences in instruction practices (e.g., McCance *et al.*, 2020). For this protocol, training is mandatory so that observers understand the codes and

how to differentiate them appropriately. The developers of the protocol indicated that 28 hours of training is needed to reach the acceptable level of inter-rater reliability between multiple observers (Hora, 2015). Also, to effectively use the protocol, it is recommended to have a meeting/interview with the instructor before observing their class. In this meeting, the goals are (1) to fill in the cover sheet of the protocol, and (2) to have information about the logistics of the class. The purpose of the cover sheet is to record the purpose of the observation, instructor characteristics, such as name, years of experience, planned goals and activities for the class, etc., and course characteristics, such as course name, department, the number of students, required or elective course, etc. The information on the logistics of the class includes recording information on the type of student seating, technology directly accessible by students and teachers, number of screens in the room and their positioning, duration of the class, recording anything unusual about the class, e.g., quiz day, etc. The basic version of the protocol provided is designed for 60 minutes of class time. It is better to have logistics information and tailor the protocol accordingly. The initial meeting also helps to decide which of the optional dimensions can be included in the observation. TDOP is best used if, besides class activities, the observer also wishes to record the type and nature of the activity. The protocol is an excellent tool to record the lecture's instructional quality (McCance *et al.*, 2020). The authors also suggested the use of TDOP online tools to record the code (Hora, 2015) and to map it on the DOLA (Differentiated Overt Learning Activities) or ICAP (Interactive, Constructive, Active and Passive) framework (Chi, 2009; Menekse *et al.*, 2013).

Classroom Observation Protocol for the Undergraduate STEM (COPUS). COPUS was designed based on four main goals: (1) to identify instructional practices, (2) to provide feedback to faculty about how the class time was spent, (3) to identify the professional development needs of the faculty, and (4) to validate the accuracy of the faculty reporting on their instructional practices (Smith *et al.*, 2013; Smith *et al.*, 2014). The protocol also had an underlying objective which evaluates the extent of traditional teaching practices in the college-level STEM courses (Smith *et al.*, 2013). This protocol used TDOP (Hora & Ferrare, 2013; Hora, 2015) as the model by utilising a similar coding structure and two-minute time interval coding patterns. On the other hand, COPUS includes a reduced number of codes (i.e., from 46 to 26) and records data in two categories as students (what students are doing, 13 codes) and instructors (what instructor is doing – 12 codes).

Strengths—The protocol is easy to follow and thus reduces the observer's judgement bias. Also, the ease of use allows it to be easily implemented by multiple observers (Lund *et al.*, 2015). Furthermore, since COPUS is designed with the professional development aspect, it can inform and support the specific needs of an instructor, as well as helping to assess students' and instructors' in-class actions efficiently and reliably (e.g., Tomkin *et al.*, 2019)). With this protocol, it is easy for the instructor to self-reflect and evaluates the quality of their instructional practices as well (Herman *et al.*, 2018).

Limitations—There are certain limitations to COPUS. First, the protocol does not account for classroom dynamics and does not determine the students' behaviours or their engagement levels. Similar to TDOP, it focuses less on students' actual work and off-task behaviours, especially in group work. It is best designed to see the frequency of response-based activities. Also, it is observed that the protocol only records students' positive behaviours (Shekhar *et al.*, 2015). The analysis of the class evaluation via this protocol results in two pie charts (one for student behaviours, one for instructor behaviours), based on the frequency of codes. However, the frequency information is limited since it does not capture the temporal information, which is essential for cross-classroom comparisons (Lund *et al.*, 2015). Finally, one of the instructor codes, 'CQ', is labeled as 'Asking a clicker question' (Smith *et al.*, 2014) and shows the bias towards clicker and respondent systems. In other words, if an instructor is using a different technology rather than clickers to actively engage students with course material during class, COPUS codes do not capture it.

Suggestions for using the protocol—The protocol is easy to use and understand and is best used to document how class time is spent. However, it is more focused on the instructor than on students. COPUS can be best used in an environment which relies on the use of clickers or classroom response systems (Lewin *et al.*, 2016). While observers can record instructional activities, the protocol does not allow us to document other types of activities besides clicker questions. Also, the protocol does not account for details related to an activity, such as how engaging that activity was. This protocol can be utilised to note the general nature of the class as passive, active, etc. Besides, similar to other protocols, COPUS requires multiple observers and training for observers before using it in real classroom settings. The developers of the protocol also created training resources for observers (Smith *et al.*, 2013).

Classroom Interactive Engagement Observation Protocol (CIEOP). CIEOP was explicitly designed for an active learning strategy called Think – Pair – Share (TPS) (Kothiyal *et al.*, 2013). In TPS, students are engaged in a three-phased in-class activity. In the first phase, students work on the task individually. In the second phase, each student is paired with another student and they discuss their individual thinking to come up with a solution. In the third phase, students share their discovered knowledge with the whole class (e.g., Lyman, 1981).

The protocol documents student engagement at each phase of the TPS activity and has specific recordable tasks associated with each stage of the TPS activity. For example, in the pair stage, the observation codes include looking around, talking off-topic with a peer and asking questions about a problem. The protocol was designed to understand the nature of students' engagement while being involved in an active learning task. The observation can be made for each student or team (Kothiyal *et al.*, 2013). The observer records the data in real-time by documenting the behaviour of one student for a specific time and then shifts to another student (Kothiyal *et al.*, 2013). In this protocol, each student can be observed at multiple time points.

Strengths—The protocol is designed to record multiple behaviours of each student (Kothiyal *et al.*, 2013) and has a set of specific practices which students may be doing

from both positive and negative directions, which help to understand the actual task being performed by an individual student. The protocol also allows the observer to record students' behaviours in small groups in large classes and provides an understanding of students' engagement (Kothiyal *et al.*, 2013).

Limitations—The protocol requires recording each student's behaviour at multiple time points, which is difficult for one observer in real classroom settings. Further, as recordings are determined based on students' facial expressions and actions, it can suffer from an observer's judgement of the behaviour. Also, the observer may not be able to document some of the behaviours while focusing on other students. In the original study, the authors used a retrospective Likert scale-based survey filled in by the students, registering their engagement (Kothiyal *et al.*, 2013). It is therefore probable that data from multiple standpoints is required to evaluate the student's engagement level (e.g., triangulation with pre- and postsurveys of engagement and students' behaviours).

Suggestions for using the protocol—The protocol is best suited to observe students' behaviours, performed in a TPS activity. The protocol can also be used to record students' behaviours in other small group activities. To use the protocol, the observer must know the different stages of the TPS activity. The protocol is easy to use if the observer focuses on a few students and records their behaviours during each phase. The authors of the protocol used one observer for ten students and used multiple observers to observe a large group of students (Kothiyal *et al.*, 2013). The protocol may require some tailoring based on the course, as besides general behaviours, there may be some course-specific behaviours that could be added, e.g., if the purpose of observation is to observe students' work when they are asked to draw or create some artefacts, the observer can record the course-specific behaviour as students showing active behaviour. Also, it is vital to decide the place or region for making an observation; the authors of the protocol used the middle rows to conduct observation as the classroom-style was V-shaped (Kothiyal *et al.*, 2013).

Student Resistance and Instructional Practices (StRIP). The StRIP protocol was developed by Shekhar *et al.* (2015) to explore the use of active learning strategies in classrooms. The primary purpose of the protocol is to understand how students respond and behave towards the instructor's instructional practices which put students in active roles. The protocol records student reactions to the use of active learning strategies (Finelli *et al.*, 2014). Also, the protocol documents specific strategies used by instructors to reduce student resistance. The authors categorised students' responses to active learning activities as both positive and negative (resistance). The students can show adverse reactions through passive behaviour – not doing the activity or partial compliance – by completing some part of the task without enthusiasm, or through open resistance, or by complaining about activities openly (Shekhar *et al.*, 2015).

This observation protocol was created using a systematic approach to gain an understanding of students' resistance to active learning and records the approximate percentage of students showing a particular type of negative response at each active learning activity. As the protocol is designed systematically, it helps to record all

details related to each active learning activity. These details include the type of active learning strategy, degree of faculty participation as high, medium or low, instructor's response during an active learning activity which also helps to comprehend the instructor's inability to implement an activity, level of students' engagement in high, medium or low categories, and students' resistance to adopt a new approach and show adverse or negative reactions. The protocol also allows observers to collect data on classroom layout and seating arrangements (Shekhar *et al.*, 2015).

Strengths—The protocol was designed with a specific goal and a systematic approach to help researchers to record instances which are otherwise not shared by participants. Further, this protocol can determine whether the instructor was able to implement active learning effectively or not. It also focuses on the kind of active learning activities done in the class. For example, the classification can be based on the ICAP (Interactive, Constructive, Active, and Passive) framework (Chi, 2009; Menekse *et al.*, 2012).

Limitations—There are several limitations. The protocol relies on recording students' engagement and resistance based on the approximate percentages of students showing that behaviour, which could be challenging to calculate in real-time and can cause reliability issues with multiple observers. The protocol also records individual student responses, ignoring the team-based response to active learning. Furthermore, in large classrooms, the recording of student resistance can only be documented if the students are open about their concerns. The protocol is additionally very open-ended, which requires a judgement from the observer. It is highly dependent on instances of active learning but provides limited information on what is accounted for as active learning. One goal of the protocol is to record students' behaviours of engagement and resistance, and a further limitation is that it may be affected by the factors of class size, classroom layout, group or individual activities, or instructor's experience, none of which are recorded by the StRIP protocol. Besides, the protocol records the difficulty level of the material covered in class, which requires either the knowledge of the content or specific cues by the instructor (Shekhar *et al.*, 2015).

Suggestions for using the protocol—The StRIP protocol is best suited to understand which of the instructors used active learning strategies. Also, the protocol helps to record the corresponding students' behaviours for these active learning strategies. The observation data helps to understand the pros and cons of each active learning activity. To use the StRIP protocol effectively, observer(s) need to roam the classrooms while observing. It would also be valuable for observers to know beforehand which learning activities will be used in a particular lecture. The authors used the protocol in a setting where active learning activities were defined and were conducted in pairs or triads (Shekhar *et al.*, 2015). It is further suggested that multiple observers are used, especially in large classes (Shekhar *et al.*, 2015), as protocol requires recording of students' engagement and resistance in real-time.

Overview and Comparisons of Eight Observation Protocols. Table 3 provides an overview and comparison of each protocol discussed in the above section. The table

Table 3. Overview of the eight protocols reviewed in this study

	RTOP	OTOP	VOS	CLOP	TDOP	COPUS	CIEOP	StRIP
Development Year	1998	2003	2003	2007	2008	2013	2013	2014
Primary Purpose	To characterise the classroom on a quantitative scale of reform	To document the impact of reform-based professional development	To record the instructional differences in the classroom	To evaluate elements of cooperative learning and teaming skills used by students	To record descriptive accounts of teaching in multidimensional terms	To record how faculty and students spend their time in the classroom	To observe student behaviours during a think-pair-share activity	To record student resistance during class activities, especially active learning activities
Focus	Effectiveness of teaching	Assess standard-based instructional practices	Nature and quality of faculty interaction, student engagement	Cooperative learning skills in engaged students	Classroom planning, dynamics, teaching	Instructor practices and students' behaviour	Students facial expression and activities	Negative resistance aspects by students
Theory, framework, or model	Constructivism	Used RTOP as model	How people learn	Cooperative learning Uses CEPT-Core as the basis	The instructional system of practice	Uses TDOP as a model	Not mentioned	Uses existing protocols for baseline
Dimensions	Lesson design and implementation, content, and classroom culture	Habits of mind, metacognition, Student discourse, challenging ideas, misconceptions, conceptual thinking, divergent thinking, interdisciplinary connections, pedagogical content knowledge, multiple representations of concepts	Classroom interaction observation (CIO), student engagement observation (SEO), Narrative notes (NN), Global ratings (GR)	**P**ositive interdependence, **I**ndividual accountability, **G**roup processing, **F**ace to face-interaction (PIG-face)	Instructional methods, pedagogical strategy, type of interaction, type of cognitive demand, use of technology	What are students doing? And what is instructor doing?	Record each student or teamwork during three phases. Think, Pair, and Share	Type of material, active learning, the degree of instructor participation, and student engagement and resistance, activity description
Student focus	No	Yes	Yes	Yes	Yes	Partially	Yes	Yes
Teacher focus	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Classroom focus	Yes	No	Partially	Yes	Yes	Partially	No	Yes
Requires training	Yes (rigorous)	Yes	Yes	Not mentioned	Yes	Yes (short)	Yes	Yes (short)

Table 3. (Continued)

	RTOP	OTOP	VOS	CLOP	TDOP	COPUS	CIEOP	StRIP
Structure of recording	Fieldnotes-based rating. Likert scale- style rating of 25 items	Ancetodal fieldnotes-based rating after the class	Video recording 3 min CIO, 30–60 sec SEO, 1–2 min NN, after class GR	5 min interval	2 min interval	2 min interval	Student by student	Each instance of active learning
Nature of Protocol	Holistic	Holistic	Segmented	Segmented	Segmented	Segmented	Holistic	Holistic
Usage reported in self-studies	153 classrooms	Three institutes, 48 sets of fieldnotes	30 courses and two institutes	Graduate engineering course	By over 300 researchers	Campus-wide 58 courses	Five activities and performed in two batches	Two institutes, four courses
Used in STEM Disciplines	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

includes multiple dimensions: the year of publication, purpose, the primary focus, underlying theory or framework, dimensions described in the protocol, the focal point of the protocols such as student-focused or teacher-focused training requirement(s), the structure of recording, how to use it in self-studies, and use of the protocol in STEM disciplines.

Please refer to Appendix 2 for the overview of studies that used protocols. Appendix 2 includes the information on sample size, focus of observation, the context of each study, training mechanism employed to use the protocol, and supplementary data sources used in conjunction with the protocols.

Validity measures of the protocols

We used the assessment triangle approach (NRC, 2001) to assess the validity of each protocol. Prior studies describe the three major aspects to evaluate and assess instruments (NRC, 2001). These aspects are: (1) cognition, referring to the theory of understanding for learning in the content domain, (2) observation, referring to what tasks are performed on a set of principles, and (3) interpretation, referring to methods and tools for interpretation of the instrument.

An assessment triangle is a model which establishes the connections between different components of an assessment system. This system focuses on assessment activities: **observation**, aligning them with knowledge and the cognitive process; **cognition**, which one wants to achieve through the measurable process; and **interpretation** (NRC, 2001).

The assessment triangle system helps to achieve a holistic and meaningful assessment and was suggested by Douglas & Purzer (2015) to validate instruments. This approach emphasises consideration of both the psychometric properties and the observation and cognition of an instrument. Furthermore, having a common mechanism to evaluate and assess an instrument gives the advantage of knowing the issues from all known aspects. We applied the suggested use of the assessment triangle to evaluate the observation protocols and used all three aspects with slight modification.

In this study, the cognition corner of the assessment triangle describes the underlying theory or framework used to design the protocol. Additionally, the cognition corner also indicates the observation aspect of the protocol and describes how that aspect of observation is recorded. The observation corner also describes the pattern of recording the information. The interpretation corner includes psychometric evidence or other validity constructs which were reported in the original study of an instrument to explain the validity and reliability of the protocol. Table 4 shows the assessment triangle aspects of all the protocols.

We recorded the underlying theory or framework used to design each observation protocol instrument in the cognition corner. It is important to recall that having a theoretical base is essential for the structured nature of an observation protocol to inform about the key aspects and direction of data interpretation. In the absence of such a basis, it is difficult to specify the instrument's domain and scope (Douglas & Purzer, 2015). Thus, the cognition corner helped in two ways: (1) describing the theoretical foundation, and (2) highlighting the observation aspect of the protocol. Among the protocols, while RTOP, OTOP, VOS, CLOP, and TDOP described the theoretical

foundations, that information was presented comprehensively in the COPUS, CIEOP, and StRIP protocols, in that the authors of these three protocols underlined the need and importance of their respective protocols, based on the literature-based evaluation of existing protocols.

Table 4 also included the pattern of recording in the observation corner of the assessment triangle. There are two categories of recording structure: (1) time based, and (2) on the occurrence of a particular event. The VOS, CLOP, TDOP and COPUS protocols are time-stamped, based on a two- to five-minute interval. The CIEOP and StRIP record information based on a particular instance. Furthermore, some protocols rely on some other mechanisms of observation to triangulate the results; for example, RTOP and OTOP are based on the holistic view and rely on field notes, while VOS requires a video recording of the classroom.

We reported psychometric or validity evidence of each protocol in the interpretation corner. Each observation protocol was evaluated using more than one validity construct, but the most common reported evidence was inter-rater reliability, either using percent agreement or Cohen-kappa values. It is however important to note that having a high coefficient alpha is not a measure of validity; instead it is a measure of internal consistency. The instrument can deviate and measure something which differs from what it is intended for (Douglas & Purzer, 2015). This deviation and measurement perspective is unclear in all protocols except CLOP and COPUS, which reported both percent agreement and Cohen Kappa reliability scores. For example, the RTOP protocol provided the estimate of reliability using best-fit regression analysis, and a point estimate of the prediction. This value may not show reliability as we are not sure how close this value is to the actual results. Using this method, a more reliable estimate would be based on an interval around the point estimate. In OTOP and VOS protocols, the authors provided a range of inter-rater reliability.

In CIEOP the authors also presented the value of inter-rater reliability. However, in all three protocols, it is unclear which metric has been used to calculate the reliability. Some protocols, for example, TDOP and StRIP protocols, did not report the value but only shared the method of calculating reliability.

All protocols except CLOP and CIEOP did some form of content validity using experts, focus groups, or other researchers' comments. However, CLOP was using the elements of cooperative learning as the foundation of the protocol. CIEOP used student response-based validation mechanisms while piloting the instrument which was used to classify student behaviour and thus provided evidence other than inter-rater reliability.

Overall, we evaluated each protocol based on the parameters of theory to support the dimensions of recording, used a defined observation method, and described the validity and reliability using appropriate measures. We found that each protocol was lacking in determining one corner of the assessment triangle. All protocols require training, and especially three of them (i.e., VOS, OTOP, and RTOP) also require an additional source of data (e.g., video recording or field notes). These missing aspects in the description of each protocol could be a reason for making these protocols challenging to understand and use.

Table 4. Assessment triangle-based evaluation of each protocol

	Cognition	Observation	Interpretation
RTOP	• Uses the constructivism approach • Dimensions of problem-solving, reasoning, communication, connections, and representation	• Compare tradition vs. reformed teaching • 25 items divided into three subsets • Likert scale style • Holistic rating of the entire class period	• 0.954 estimate of reliability using best-fit regression modeling • Face, predictive and construct validity for each subscale • Exploratory factor analysis to uncover underlying relationships
OTOP	• Use the RTOP and 5E model as the base • Characterises lesson on ten possible indicators of standard-based instructional practices	• Narrative fieldnotes • Likert scale style rating of 10 indicators	• Content validity with both faculty and graduate students • Percent agreement of each practice ranged 57–100%
VOS	• Uses 'How People Learn' • Uses four methods of data collection: Student-teacher interaction, student academic engagement, narrative notes, ratings of effective teaching	• Time-stamped 3 min CIO, 30–60 sec SEO, 1–2 min NN, and after class GR • Real-time observation • Compare the HPL-based course with traditional course	• Inter-rater reliability as 85–91% • Content validity using 11 content experts • Assess the presence of HPL-oriented classroom activities
CLOP	• Use 'Cooperative learning.' • Determine the engagement level of students on essential elements of cooperative learning	• 5 min interval • A recorded instance of each element where student engagement is recorded as low, medium, high or not observed	• Agreement of four ratings as 75–95% • Cohen Kappa for inter-rater reliability agreement = 0.67
TDOP	• Uses an instructional system of practices • Dimensions of classroom dynamics, interactions among actors, and artifacts • Trained observers	• 2 min time interval • 46 Codes based (used 33 codes in the self-study) • Determined active learning modalities (active, constructive, interactive, and problem-solving)	• Content validity through the educational researcher • Inter-coder reliability with the use of 2% of dataset and Cohen's Kappa • Used code frequencies, and social network analysis techniques to depict data

Table 4. (Continued)

	Cognition	Observation	Interpretation
COPUS	• Classroom Behaviours • What students are doing • What instructor is doing • Uses TDOP as a model	• 2 min interval • 12 instructor codes • 13 Student codes	• Validity was based on experts and teachers' feedback • Jaccard similarity score for codes > 0.9 • Cohen's Kappa for observer reliability 0.79–0.87 • Descriptive and chart based analysis
CIEOP	• Student engagement of what they are doing • Student Behaviours • Specific to activity: Think-Pair-Share	• Each student observed for a certain amount of time • Multiple observers • Six to 11 cycles • Real-time observation	• Validated using students' responses and pilot • 90% intercoder reliability • Triangulated with survey data • Classified behaviours and engagements
StRIP	• Students' response to active learning • Determined teacher participation and type of student resistance	• Records at each specific instance of active learning • Two-page protocol • Open-ended recording • Relies on video recording as well • Two variations: First day, and daily instrument	• Validated using focus group with engineering students using images of observed classrooms • Also validated using experts' feedback • Inter-rater reliability was reported as high • Descriptive and classification approach to describe data

Discussion

Researchers often face the challenge of selecting an observation protocol that is best suited for their research study. The primary goal of this study was to provide an overview of eight observation protocols that have been used to evaluate classroom practices, instructional methods, or students' engagement and behaviours. For each observation protocol, we focused on answering each protocol's primary purpose, strengths, limitations and validity.

Each protocol included in this review was designed with a unique focus. For instance, some protocols were designed with instructor-focused (Piburn & Sawada, 2000) or student-focused dimensions (Kothiyal *et al.*, 2013), with few protocols focused on both dimensions (Hora & Ferrare, 2013; Smith *et al.*, 2013). It is noteworthy that four protocols (VOS, TDOP, COPUS and StRIP) have aspects of observing lectures from the perspective of all three actors (instructor, student and classroom dynamics). While TDOP and StRIP give similar importance to all three actors, both VOS and COPUS are more instructor-focused.

The context of the protocols varied from the emphasis on a particular grade level or discipline. For example, RTOP was initially designed in the context of observing K-12 teachers on the fidelity of implementation related to reformed practices with K-12 classes. On the other hand, some protocols were explicitly designed for undergraduate courses, such as OTOP, VOS and TDOP. OTOP was designed to document the impact of reform-based professional development training of an undergraduate faculty on their instructional methods and student behaviours. VOS was initially designed with a specific focus on biomedical engineering classrooms and is useful in observing teaching and learning experiences (Gazca *et al.*, 2009).

Similarly, TDOP was developed to document the instructor and student interactions in college classrooms. In addition to contexts related to grade level, some protocols were explicitly designed to record particular activity details. For example, CLOP is very specific to group activities, while COPUS focuses on active learning in classrooms and aims to document students' reactions to these activities. Similarly, CIEOP is very specific to one active learning activity of TPS. StRIP emphasised recording of both positive and negative student reactions to active learning methods.

Based on the goal of the protocol, the design strategy could focus on evaluating classroom activities from the perspective of what is happening in class (Hora & Ferrare, 2013), or on the fidelity of implementation of a particular intervention (Kothiyal *et al.*, 2013). Some were also designed to study student behaviours and engagement (Kern *et al.*, 2007; Cox & Cordray, 2008). We identified that these protocols could be classified into three distinct categories which include: (1) protocols for evaluation and feedback for instructional methods, (2) protocols for evaluating pedagogical strategies, and (3) protocols for students' engagement or behaviour.

RTOP, OTOP, TDOP and COPUS are the most notable protocols for evaluating a faculty and providing them feedback on their teaching approaches. RTOP is being used as a tool for course evaluation (e.g., Amrein-Beardsley & Popp, 2012; Helpert & Husman, 2017; Addy & Blanchard, 2010); OTOP is best suited to evaluate standard-based instruction and the impact of reform-based professional development; TDOP is best suited for providing formative feedback for the improvement of instructional

methods; and COPUS is most suited to determine faculty practices and, accordingly, professional development needs of the faculty (e.g., Herman *et al.*, 2018).

Among the protocols for evaluating pedagogical strategies, we found that VOS, TDOP and StRIP are the best suited observation protocols. While VOS and TDOP help to identify the type of instructional strategy used in class, StRIP is used to account for active learning strategies. Although TDOP and COPUS offer opportunities to record students' behaviours, their focus is on how students spend their time in a broader spectrum and thus fail to catch students' engagement instances.

For the third category, we found that OTOP, VOS, CLOP, CIEOP and StRIP protocols are suited to document students' engagement. These protocols have a different goal to observe students' engagement. OTOP student behaviour rating includes students' engagement with lesson content, peers and task. VOS requires a real-time headcount of engaged students. CLOP and CIEOP both evaluate student engagement in group tasks, but CIEOP is very specific to the TPS activity. Alternatively, StRIP allows observers to record students' negative resistance while being engaged in a task. It is also noted that StRIP is the only protocol that covered the information about the type of material involved and student engagement.

We also evaluated all protocols on a common mechanism of assessment triangle, which helped us review each protocol's authenticity and validity. Researchers should use the appropriate protocol based on their need for recording and utilising proper measures to reduce observer bias and improve accuracy. One proposed method to avoid these biases is the observer's training in an appropriate setting. Furthermore, the establishment of inter-rater reliability among observers is essential.

The use of structured observation protocols reduces some of the known limitations of classroom observations. For example, structured observation protocols help to reduce the observer's preferences bias while recording information and collecting data. Also, the observation protocols allow the studying of multiple aspects of the same phenomenon which may not be feasible otherwise. On the other hand, observation protocols have some limitations. First, the observer's presence or the reactivity effect on the observed may be a limitation. The reactivity effect can result in escalation or de-escalation of the normal behaviour of the observed (aka. Hawthorne effect) (Mayo *et al.*, 1939; Turnock & Gibson, 2001). Second, the observation process is time-consuming. Third, the observer's accuracy may be compromised due to the understanding of the behaviour or appropriately recording the response according to the protocol's guidelines.

Limitations

In this systematic review, we followed a repeatable process of selection and inclusion. However, there are limitations of the review studies, including quality constraints, selection bias and publication bias. For example, we have not excluded studies based on their quality and reporting mechanisms (Slavin, 1984). We focused on the selection of authentic databases, but the actual quality of the study was not the focal point of this review. Additionally, the authors made a judgement call for the appropriate

databases, which may have limited the selection of studies. This call may introduce bias in the selection mechanism, and studies that are published at other venues may not have been part of this study. Although we did not favour the positive results-based studies only (Rothstein, Sutton, & Borenstein, 2006), our results are only based on exemplary studies for each protocol (Borrego *et al.*, 2015)

Conclusion

The primary purpose of this systematic review was to provide a comparative overview of commonly used structured observation protocols in postsecondary STEM disciplines systematically. We outlined these protocols and provided the characteristics of eight of them. Also, we believe that this literature review will help future researchers in deciding the appropriate tool for their research purpose. Based on the above discussion, we also have suggestions for future researchers, especially new protocol developers. For example, they should state the precise reasons for developing a new observation protocol. Besides focusing on reporting agreement and reliability, the protocols should focus on the theoretical foundations for the protocol. One common mechanism future developers can utilise is to make use of the assessment triangle as a model and describe the three corners in detail. The protocol developers should also explain the metric used for calculating the psychometric properties of an instrument, and the emphasis should be on both the internal consistency of the items and the overall multiple observers' agreement.

We noticed that a common limitation among the existing studies which describe the protocols was detailing how the observers can be trained to conduct the observation. Although some protocols provided a training manual, these details were not mentioned in the studies published for introducing the protocol. Also, we suggest that future developers should describe exemplary research questions for the protocol and describe the ideal learning and teaching settings to use their tools. It is also important that the introductory documents or the observation protocol manuals should include the appendix of sample data and how it has been transformed to make the observation-related claims.

Overall, this study provides an overview and guidelines for using commonly used observation protocols in college classrooms. This study provides a point of reference to other researchers for knowing the details of these protocols and helping them while making decisions on what observation protocols to use, based on their goals, research questions, and research design.

Conflict of interest

The authors certify that they have no potential conflict of interest.

Data availability statement

This paper uses existing published articles as the data set. The list of those papers is available in Appendix 1.

References

Adamson, S.L., Banks, D., Burtch, M., Cox, F. III, Judson, E., Turley, J.B. *et al.* (2003) Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement, *Journal of Research in Science Teaching*, 40(10), 939–957.

Addy, T.M. & Blanchard, M.R. (2010) The Problem with Reform from the Bottom up: Instructional practises and teacher beliefs of graduate teaching assistants following a reform-minded university teacher certificate programme, *International Journal of Science Education*, 32(8), 1045–1071.

American Association for the Advancement of Science. (2013) *Describing and measuring undergardaute STEM teaching practices*. Available online at: http://www.nsf-i3.org/resources/view/describing_and_measuring_teaching_practices/ (accessed 20 August 2020).

Amrein-Beardsley, A. & Popp, S.E.O. (2012) Peer observations among faculty in a college of education: Investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP), *Educational Assessment, Evaluation and Accountability*, 24(1), 5–24.

American Society for Engineering Education (ASEE). (2012) *Innovation with impact: Creating a culture for scholarly and systematic innovation in engineering education* (Washington, DC, American Society for Engineering Education (ASEE)).

Banchi, H.M. (2009) *Learning from the best: Overcoming barriers to reforms-based elementary science teaching* (ERIC).

Berk, R.A. (2005) Survey of 12 strategies to measure teaching effectiveness, *International Journal of Teaching and Learning in Higher Education*, 17(1), 48–62.

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

Brophy, J. & Good, T. (1986) *Teacher-effects results. Handbook of research on teaching* (New York, Macmillan).

Borrego, M., Foster, M.J. & Froyd, J.E. (2015) What is the state of the art of systematic review in engineering education?, *Journal of Engineering Education*, 104(2), 212–242.

Borrego, M. & Henderson, C. (2014) Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies, *Journal of Engineering Education*, 103(2), 220–252.

Campbell, T., Abd-Hamid, N.H. & Chapman, H. (2010) Development of instruments to assess teacher and student perceptions of inquiry experiences in science classrooms, *Journal of Science Teacher Education*, 21(1), 13–30.

Chi, M.T.H. (2009) Active-constructive-interactive: A conceptual framework for differentiating learning activities, *Topics in Cognitive Science*, 1(1), 73–105.

Cox, M.F. & Cordray, D.S. (2008) Assessing Pedagogy in bioengineering classrooms: Quantifying elements of the "How People Learn" model using the VaNTH observation system (VOS, *Journal of Engineering Education*, 97(4), 413–431.

Cox, M.F., Hahn, J., McNeill, N., Cekic, O., Zhu, J. & London, J. (2011) Enhancing the quality of engineering graduate teaching assistants through multidimensional feedback, *Advances in Engineering Education*, 2(3), n3.

Douglas, K.A. & Purzer, Ş. (2015) Validity: Meaning and relevancy in assessment for engineering education research, *Journal of Engineering Education*, 104(2), 108–118.

Ebert-May, D., Derting, T.L., Henkel, T.P., Middlemis Maher, J., Momsen, J.L., Arnold, B. *et al.* (2015) Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development, *CBE—Life Sciences Education*, 14(2), ar22.

Ebert-May, D., Derting, T.L., Hodder, J., Momsen, J.L., Long, T.M. & Jardeleza, S.E. (2011) What we say is not what we do: effective evaluation of faculty professional development programs, *BioScience*, 61(7), 550–558.

Finelli, C.J., Daly, S.R. & Richardson, K.M. (2014) Bridging the research-to-practice gap: Designing an institutional change plan using local evidence, *Journal of Engineering Education*, 103(2), 331–361.

Finelli, C.J., DeMonbron, M., Shekhar, P., Borrego, M., Henderson, C., Prince, M. *et al.* (2014) A classroom observation instrument to assess student response to active learning. in: *Proceedings of IEEE Frontiers in Education Conference (FIE)* (IEEE), 1–4.

Frost, L., Goodson, L., Greene, J., Huffman, T., Kunberger, T. & Johnson, B. (2018) SPARCT: A STEM professional academy to reinvigorate the culture of teaching, *Journal of STEM Education*, 19(1), 62–69.

Gazca, L., Palou, E., Lopez-Malo, A. & Garibay, J.M. (2009) Capturing differences of engineering design learning environments by means of VaNTH observation system. in: *Proceedings of the 39th ASEE/IEEE Frontiers in Education Conference*.

Harris, A.H. & Cox, M.F. (2003) Developing an observation system to capture instructional differences in engineering classrooms, *Journal of Engineering Education*, 92(4), 329–336.

Henry, M.A., 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, *Document Number*.

Herman, G.L., Greene, J.C., Hahn, L.D., Mestre, J.P., Tomkin, J.H. & West, M. (2018) Changing the teaching culture in introductory STEM courses at a large research university, *Journal of College Science Teaching*, 47(6), 32–38.

Hilpert, J.C. & Husman, J. (2017) Instructional improvement and student engagement in post-secondary engineering courses: The complexity underlying small effect sizes, *Educational Psychology*, 37(2), 157–172.

Hora, M.T. (2015) Toward a descriptive science of teaching: How the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms, *Science Education*, 99(5), 783–818.

Hora, M.T., Bouwma-Gearhart, J. & Park, H.J. (2017) Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning, *The Review of Higher Education*, 40 (3), 391–426.

Hora, M.T. & Ferrare, J.J. (2013) Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching, *Journal of the Learning Sciences*, 22(2), 212–257.

Huntley, M.A. (2009) Measuring curriculum implementation, *Journal for Research in Mathematics Education*, 40(4), 355–362.

Johnson, D.W. & Johnson, R.T. (1999) Making cooperative learning work, *Theory into Practice*, 38 (2), 67–73.

Johnson, D.W., Johnson, R.T. & Smith, K.A. (1998) Cooperative learning returns to college what evidence is there that it works?, *Change: The Magazine of Higher Learning*, 30(4), 26–35.

Kern, A.L., Moore, T.J. & Akillioglu, F.C. (2007) Cooperative learning: Developing an observation instrument for student interactions. in: *Proceedings of 37th Annual Frontiers Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, T1D–1* (IEEE).

Kothiyal, A., Majumdar, R., Murthy, S. & Iyer, S. (2013) Effect of think-pair-share in a large CS1 class: 83% sustained engagement. in: *Proceedings of the ninth annual international ACM conference on International computing education research*, 137–144. https://doi.org/10.1145/2493394.2493408

Lawrenz, F., Huffman, D., Appeldoorn, K. & Sun, T. (2002) *CETP core evaluation, classroom observation handbook* (Minneapolis, MN, CAREI).

Lawson, A., Benford, R., Bloom, I. & Carlson, M. (2002) Evaluating college science and mathematics instruction, *Journal of College Science Teaching*, 31(6), 388.

Lewin, J.D., Vinson, E.L., Stetzer, M.R. & Smith, M.K. (2016) A campus-wide investigation of clicker implementation: The status of peer discussion in STEM classes, *CBE—Life Sciences Education*, 15(1), ar6.

Liang, L.L., Fulmer, G.W., Majerich, D.M., Clevenstine, R. & Howanski, R. (2012) The effects of a model-based physics curriculum program with a physics first approach: A causal-comparative study, *Journal of Science Education and Technology*, 21(1), 114–124.

Lund, T.J., Pilarz, M., Velasco, J.B., Chakraverty, D., Rosploch, K., Undersander, M. *et al.* (2015) The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice, *CBE—Life Sciences Education*, 14(2), ar18.

Luo, Y. (2017) Design fixation and cooperative learning in elementary engineering design project: A case study, *International Electronic Journal of Elementary Education*, 8(1), 133–146.

Lyman, F. (1981) The responsive classroom discussion, in: A.S. Anderson (Ed) *Mainstreaming digest* (College Park, MD, University of Maryland College of Education).

Maresca, P., Guercio, A., Stanganelli, L. & Arndt, T. (2014) Experiences in collaborative learning, *Journal of E-Learning and Knowledge Society*, 10(3), 121–145.

Mayo, E., Roethlisberger, F. & Dickson, W. (1939) *Management and the worker* (Cambridge, MA, Harvard University Press).

McCance, K., Weston, T. & Niemeyer, E. (2020) Classroom observations to characterize active learning within introductory undergraduate science courses breadcrumb, *Journal of College Science Teaching*, 49(4). https://www.nsta.org/journal-college-science-teaching/journal-college-science-teaching-marchapril-2020/classroom.

Menekse, M., Chi, M.T.H., Baker, D. & Middleton, J. (2012) *Interactive-Constructive-Active-Passive: The Relative Effectiveness of Differentiated Activities on Students' Learning.*

Menekse, M., Stump, G.S., Krause, S. & Chi, M.T.H. (2013) Differentiated overt learning activities for effective instruction in engineering classrooms, *Journal of Engineering Education*, 102(3), 346–374.

Middleton, J.A., Krause, S., Beeley, K., Judson, E., Ernzen, J. & Culbertson, R. (2015, October) Examining the relationship between faculty teaching practice and interconnectivity in a social network. in: *2015 IEEE Frontiers in Education Conference (FIE)* (IEEE), 1–7.

Millis, B.J. (1992) Conducting effective peer classroom observations, *To Improve the Academy*, 11 (1), 189–206.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & The PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement, *PLOS Medicine*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Morrell, P.D., Wainwright, C. & Flick, L. (2004) Reform teaching strategies used by student teachers, *School Science and Mathematics*, 104(5), 199–213.

Moore, M., (2017) *Assessing Student Behaviors and Motivation for Actively Learning Biology. (Doctoral Dissertation)*. Oklahoma: Oklahoma State University.

Mulhall, A. (2003) In the field: Notes on observation in qualitative research, *Journal of Advanced Nursing*, 41(3), 306–313.

NRC (2001) *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Parahoo, K. (2014) *Nursing research: Principles, process and issues* (Macmillan International Higher Education).

Pianta, R.C. & Hamre, B.K. (2009) Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity, *Educational Researcher*, 38(2), 109–119.

Piburn, M. & Sawada, D. (2000) *Reformed Teaching Observation Protocol (RTOP) Reference Manual*. Technical, Report.

Pretzlik, U. (1994) Observational methods and strategies, *Nurse Researcher*, 2(2), 13–21.

President's Council of Advisors on Science and Technology (PCAST) (2012) *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics* (Washington DC, President's Council of Advisors on Science and Technology (PCAST))

Rivera, N. (2013) *Cooperative learning in a community college setting: Developmental coursework in mathematics* (Arizona State University).

Ross, S.M., Smith, L.J., Alberg, M. & Lowther, D. (2004). Using classroom observation as a research and formative evaluation tool in educational reform. in: *Observational research in US classrooms: New approaches for understanding cultural and linguistic diversity*, 144–173.

Rothstein, H.R., Sutton, A.J. & Borenstein, M. (Eds.), (2006). *Publication bias in metaanalysis: Prevention, assessment and adjustments*. Chichester, UK: John Wiley & Sons

Rushton, G.T., Lotter, C. & Singer, J. (2011) Chemistry teachers' emerging expertise in inquiry teaching: The effect of a professional development model on beliefs and practice, *Journal of Science Teacher Education*, 22(1), 23–52.

Sawada, D., Piburn, M.D., Judson, E., Turley, J., Falconer, K., Benford, R. *et al.* (2002) Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol, *School Science and Mathematics*, 102(6), 245–253.

Sawada, D., Piburn, M., Turley, J., Falconer, K., Benford, R., Bloom, I. *et al.* (2000) *Reformed teaching observation protocol (RTOP) training guide*. ACEPT IN-002. Arizona Board of Regents.

Shekhar, P., Demonbrun, M., Borrego, M., Finelli, C., Prince, M., Henderson, C. *et al.* (2015) Development of an observation protocol to study undergraduate engineering student resistance to active learning, *International Journal of Engineering Education*, 31(2), 597–609.

Slavin, R.E. (1984) Meta-analysis in education: How has it been used? *Educational Researcher*, 13 (8), 6–15. https://doi.org/10.3102/0013189X013008006

Smith, M.K., Jones, F.H.M., Gilbert, S.L. & Wieman, C.E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE – Life Sciences Education*, 12(4), 618.

Smith, L., Martin, W.G., Wan, A., Duenas, G., Shumack, K. & Beziat, T.L. (2015) Using reform pedagogy to improve students'application skills in college remedial mathematics courses, *Mathematics and Computer Education*, 49(2), 124.

Smith, M.K., Vinson, E.L., Smith, J.A., Lewin, J.D. & Stetzer, M.R. (2014) A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE – Life Sciences Education*, 13(4), 624.

Stallings, J.A. & Mohlman, G.G. (1988) Classroom observation techniques, *Educational Research, Methodology, and Measurement: An International Handbook*, 469–474.

Teaching Dimension Observation Protocol. (2010) http://tdop.wceruw.org/(accessed 8 March 2018).

Teasdale, R., Manduca, C.A., Mcconnell, D.A., Bartley, J.K., Bruckner, M.Z., Farthing, D. *et al.* (2014) Observations of undergraduate geoscience instruction in the US: Measuring student centered teaching, *AGU Fall Meeting Abstracts*.

Teasdale, R., Viskupic, K., Bartley, J.K., McConnell, D., Manduca, C., Bruckner, M. *et al.* (2017) A multidimensional assessment of reformed teaching practice in geoscience classrooms, *Geosphere*, 13(2), 608–627.

Tolentino, L., Birchfield, D., Megowan-Romanowicz, C., Johnson-Glenberg, M.C., Kelliher, A. & Martinez, C. (2009) Teaching and learning in the mixed-reality science classroom, *Journal of Science Education and Technology*, 18(6), 501–517.

Tomkin, J.H., Beilstein, S.O., Morphew, J.W. & Herman, G.L. (2019) Evidence that communities of practice are associated with active learning in large STEM lectures, *International Journal of STEM Education*, 6(1), 1.

Turnock, C. & Gibson, V. (2001) Validity in action research: a discussion on theoretical and practice issues encountered whilst using observation to collect data, *Journal of Advanced Nursing*, 36 (3), 471–477.

Wainwright, C.L., Flick, L.B. & Morrell, P.D. (2003) Development of instruments for assessment of instructional practices in standards-based teaching, *Journal of Mathematics and Science: Collaborative Explorations*, 6(1), 21–46.

Wainwright, C., Morrell, P.D., Flick, L. & Schepige, A. (2004) Observation of reform teaching in undergraduate level mathematics and science courses, *School Science and Mathematics*, 104(7), 322–335.

Waxman, H.C. & Padron, Y.N. (2004) The uses of the classroom observation schedule to improve classroom instruction. in: *Observational research in US classrooms: New approaches for understanding cultural and linguistic diversity*, 72–96.

Whiteside, A., Brooks, D.C. & Walker, J.D. (2010) Making the case for space: Three years of empirical research on learning environments, *Educause Quarterly*, 33(3), 11.

Appendix 1 . List of selected studies

Category: Articles that introduce the protocol

Author(s)' Names	Title
Protocol Name: RTOP	
Sawada, D., Piburn, M.D., Judson, E., Falconer, K. & others. (2002)	Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol.
Piburn, M.D. & Sawada, D. (2000)	Reformed Teaching Observation Protocol (RTOP): Reference Manual.
Protocol Name: OTOP	
Wainwright, C. L., Flick, L., & Morrell, P. (2003).	The Development of Instruments for Assessment of Instructional Practices in Standards-Based Teaching.
Protocol Name: VOS	
Harris, A.H., & Cox, M.F. (2003)	Developing an Observation System to Capture Instructional Differences in Engineering Classrooms.
Protocol Name: CLOP	
Kern, A.L, Moore, A.J., & Akillioglu F.C. (2007)	Cooperative Learning: Developing an Observation Instrument for Student Interaction.
Protocol Name: TDOP	
Hora, M.T. & Ferrare, J.J. (2012)	Instructional Systems of Practice: A Multidimensional Analysis of Math and Science Undergraduate Course Planning and Classroom Teaching.
Hora. M.T. (2013)	Toward a Descriptive Science of Teaching: How the TDOP Illuminates the Multidimensional Nature of Active Learning in Postsecondary Classrooms.
Protocol Name: COPUS	
Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013).	The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterise University Stem Classroom Practices
Protocol Name: CIEOP	
Kothiyal, A., Majumdar, R., Murthy, S., & Iyer, S. (2013)	Effect of Think-Pair-Share in a Large CS1 Class: 83% Sustained Engagement
Protocol Name: StRIP	
Finelli, C. J., DeMonbron, M., Shekhar, P., Borrego, M., Henderson, C., Prince, M., & Waters, C. K. (2014)	A Classroom Observation Instrument to Assess Student Response to Active Learning.

Appendix 1. (Continued)

Author(s)' Names	Title
Shekhar, P., Demonbrun, M., Borrego, M., Finelli, C., Prince, M., Henderson, C., & Waters, C. (2015)	Development of an Observation Protocol to Study Undergraduate Engineering Student Resistance to Active Learning.

Category: Articles that use the protocol

Author(s)' Names	Title
Protocol Name: RTOP	
Frost, L., Goodson, L., Greene, J., Huffman, T., Kunberger, T., & Johnson, B. (2018)	SPARCT: A STEM Professional Academy to Reinvigorate the Culture of Teaching
Smith, L., Martin, W. G., Wan, A., Duenas, G., Shumack, K., & Beziat, T. L. (2015)	Using Reform Pedagogy to Improve Students' application Skills in College Remedial Mathematics Courses
Reinbold, S. L. (2018)	Critical Thinking Assessment of Students in Nonmajors Biology Classes with Corn or Fly Genetics Laboratory Studies.
Diersen, G. T. (2011).	Team Echinacea & Construction of a Key Using Online Images of Fresh Prairie Plant Pollen.
Middleton, J. A., Krause, S., Beeley, K., Judson, E., Ernzen, J., & Culbertson, R. (2015)	Examining the Relationship between Faculty Teaching Practice and Interconnectivity in a Social Network.
Hilpert, J. C., & Husman, J. (2017).	Instructional Improvement and Student Engagement in Post-Secondary Engineering Courses: The Complexity Underlying Small Effect Sizes.
Ryker, K., & McConnell, D. (2014)	Can Graduate Teaching Assistants Teach Inquiry-Based Geology Labs Effectively?
Protocol Name: OTOP	
Wainwright, C., Morrell, P. D., Flick, L., & Schepige, A. (2004)	Observation of Reform Teaching in Undergraduate Level Mathematics and Science Courses.
Protocol Name: VOS	
Cox, M.F., & Cordray, D.D. (2008)	Assessing Pedagogy in Bioengineering Classrooms: Quantifying Elements of the 'How People Learn' Model using the VaNTH Observation System (VOS).
Protocol Name: TDOP	
Hora, M.T. (2014)	Exploring Faculty Beliefs about Student Learning and Their Role in Instructional Decision-Making.

Appendix 1. . (Continued)

Author(s)' Names	Title
Oleson, A. & Hora, M.T. (2014)	Teaching the Way they were Taught? Revisiting the Sources of Teaching Knowledge and the Role of Prior Experience in Shaping Faculty Teaching Practices.
Hora, M.T. & Ferrare, J.J. (2014)	Remeasuring Postsecondary Teaching: How Singular Categories of Instruction Obscure the Multiple Dimensions of Classroom Practice.
McCance, K., Weston, T., & Niemeyer, E. (2020).	Classroom Observations to Characterise Active Learning Within Introductory Undergraduate Science Courses Breadcrumb.
Finelli, C. J., Daly, S. R., & Richardson, K. M. (2014)	Bridging the Research-to-Practice Gap: Designing an Institutional Change Plan using Local Evidence
Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017).	Data Driven Decision-Making in the Era of Accountability: Fostering Faculty Data Cultures For Learning.
Protocol Name: COPUS	
Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014)	A Campus-Wide Study of STEM Courses: New perspectives on Teaching Practices and Perceptions
Herman, G. L., Greene, J. C., Hahn, L. D., Mestre, J. P., Tomkin, J. H., & West, M. (2018).	Changing the Teaching Culture in Introductory STEM Courses at a Large Research University
Callens, M. V., Kelter, P., Motschenbacher, J., Nyachwaya, J., Ladbury, J. L., & Semanko, A. M. (2019).	Developing and Implementing a Campus-Wide Professional Development Program: Successes and Challenges
Chacón-Díaz, L. B. (2020).	An Explanatory Case Study of Behaviours, Interactions, and Engagement in an Introductory Science Active Learning Classroom (ALC).
Strubbe, L. E., Stang, J., Holland, T., Sherman, S. B., & Code, W. J. (2019)	Faculty Adoption of Active Learning Strategies via Paired Teaching: Conclusions from Two Science Departments
Holt, E. A., & Nielson, A. (2019).	Learning Communities and Unlinked Sections: A Contrast of Student Backgrounds, Student Outcomes, and In-class Experiences
Tomkin, J. H., Beilstein, S. O., Morphew, J. W., & Herman, G. L. (2019).	Evidence that Communities of Practice are Associated with Active Learning in Large STEM Lectures.
Protocol Name: CIEOP	
Reddy, P. D., Mishra, S., Ramakrishnan, G., & Murthy, S. (2015)	Thinking, Pairing, and Sharing to Improve Learning and Engagement in a Data Structures and Algorithms (DSA) Class.
Protocol Name: StRIP	
Finelli, C. J., Nguyen, K., DeMonbrun, M., Borrego, M., Prince, M., Husman, J., . . . & Waters, C. K. (2018)	Reducing student resistance to active learning: Strategies for instructors.

Appendix 2. Characteristics of articles that uses protocols

NOTE: N/A = NOT APPLICABLE

Protocol Name: RTOP

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Frost, L., Goodson, L., Greene, J., Huffman, T., Kunberger, T., & Johnson, B. (2018)	36 instructors	To monitor the impact and progress of year-long professional development	N/A	Conduct peer observation during an introductory STEM course	N/A	Video recordings
Smith, L., Martin, W. G., Wan, A., Duenas, G., Shumack, K., & Beziat, T. L. (2015)	43 students	To corroborate that two types of instruction, i.e., traditional lecture and reform- based are different	N/A	To compare the difference in application skills between two groups of students	N/A	N/A
Reinbold, S. L. (2018)	Six sections	To evaluate the instruction of a lesson and to understand the variation in instruction	N/A	To compare biology students in a non-major course where students participated in two different kinds of activities	N/A	N/A
Diersen, G. T. (2011)	N/A	To construct and guide the lesson to be inquiry-based	N/A	To evaluate pollen lesson implementation in a biology course	N/A	N/A

Appendix 2. (Continued)

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Middleton, J. A., Krause, S., Beeley, K., Judson, E., Ernzen, J., & Culbertson, R. (2015)	21 instructors	To identify specific teaching practices	N/A	To determine the degree of social connectedness in a faculty with STEM disciplines, and learner-centred practices	N/A	Fieldnotes an interviews
Hilpert, J. C., & Husman, J. (2017)	11 engineering professors after yearlong professional development (PD)	To conduct course evaluations of instructors	N/A	To examine the impact of professional development based on improved instruction	RTOP online training	N/A
Ryker, K., & McConnell, D. (2014)	48 observations	To assess Teaching practices	N/A	To examine the implementation of teaching strategies by graduate teaching assistants (GTAs) in inquiry-based course	N/A	N/A

Protocol Name: OTOP

Appendix 2. (Continued)

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Wainwright, C., Morrell, P. D., Flick, L., & Schepige, A. (2004)	Five institutes, 37	observations	To measure the	implementation of reform-based practices	N/A	What elements of reform teaching are being utilised by instructors
Conducted pair observations to ensure reliability	Interviews Fieldnotes					

Protocol Name: VOS

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Cox, M.F., & Cordray, D.D. (2008)	28 bioengineering courses	To assess the pedagogical difference between lecture-based and HPL-oriented courses	N/A	Application of HPL index that distinguishes the pedagogical style	Use VOS training manual	N/A

Appendix 2. (Continued)

Protocol Name: TDOP

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Hora, M.T. (2014)	56 faculty	Instructors' use of teaching methods and instructional technology	N/A	How do faculty beliefs about student learning influence their plans to teach courses?	Three-day training	Interviews
Oleson, A. & Hora, M.T. (2014)	53 faculties, two in-depth analyses	Explore the relationship between experimental knowledge of classroom teaching practices	N/A	Role of prior experience in the development of teaching practices	Three-day extensive training	Interviews
Hora, M.T. & Ferrare, J.J. (2014)	58 faculties	To understand teaching dimensions used by the faculty	N/A	To study post-secondary teaching that records five distinct dimensions of teaching practices	Extreme training	N/A
McCance, K., Weston, T., & Niemeyer, E. (2020).	Six courses	To characterise differences in instructional practices among undergraduate science courses	N/A	To characterise how active learning practices have been incorporate in the courses	TDOP training	N/A
Finelli, C. J., Daly, S. R., & Richardson, K. M. (2014)	30 faculties	To investigate teaching practices currently in use by faculty	N/A	Design and implementation of an effective change plan for institute and courses	N/A	Focus group discussions
Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017).	59 faculties	To understand the use of the teaching method by the instructor	N/A	How faculty uses teaching to relate data?	28 hr training	Interviews

Appendix 2. (Continued)

Protocol Name: COPUS

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014)	51 courses	To understand the current structure of STEM teaching practices	RTOP	How faculty's current practices can inform the design of professional development?	Short training	Surveys
Herman, G. L., Greene, J. C., Hahn, L. D., Mestre, J. P., Tomkin, J. H., & West, M. (2018).	49 depts	How both instructor and students spend class time?	N/A	To discuss the results of efforts to implement evidence-based instructional practices	N/A	N/A
Callens, M. V., Kelter, P., Motschenbacher, J., Nyachwaya, J., Ladbury, J. L., & Semanko, A. M. (2019).	30 STEM faculties	To observe classroom practices used by the instructor	RTOP, TDOP	To implement a campus-wide professional development program	N/A	N/A
Chacón-Díaz, L. B. (2020).	28 students, one course	To observe the implementation of active learning classes	N/A	To describe students' behaviours and interaction in an active learning class	Year-long training experience	Fieldnotes and surveys
Strubbe, L. E., Stang, J., Holland, T., Sherman, S. B., & Code, W. J. (2019)	14 faculties	To document the evolution in faculty teaching	N/A	How faculty is trained for active learning practices in a paired teaching environment	N/A	Interviews

Appendix 2. (Continued)

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Holt, E. A., & Nielson, A. (2019).	13 students	To observe common behaviour and teaching practices	N/A	To compare participated students' academic performance, retention, student background and engagement	N/A	Surveys
Tomkin, J. H., Beilstein, S. O., Morphew, J. W., & Herman, G. L. (2019).	25 courses	To compare teaching practices of faculty who participate in communities of practice with those who did not	N/A	To explore whether participation in communities of practice is related to active learning	N/A	N/A

Protocol Name: CIEOP

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Reddy, P. D., Mishra, S., Ramakrishnan, G., & Murthy, S. (2015)	12 lecturers, 90 students	To record students' behaviours during think-pair-share activities	N/A	To investigate the effect of think-pair-share activity on students' conceptual understanding, engagement, and perception of TPS in the course	N/A	Surveys Interviews

Appendix 2. (Continued)

Protocol Name: StRIP

Author(s)' Names	Sample size	Focus of observation	Other considered protocols	Context of the study	Training mechanism	Supplementary data sources
Finelli, C. J., Nguyen, K., DeMonbrun, M., Borrego, M., Prince, M., Husman, J.,. . . & Waters, C. K. (2018)	18 faculties, 1051 students	To observe students' response to active learning	N/A	To understand barriers of student response and instructional strategies	N/A	Interviews