Williams et al. International Journal of STEM Education (2015) 2:18
DOI 10.1186/s40594-015-0031-y

International Journal of STEM Education
a SpringerOpen Journal

RESEARCH Open Access

Describing undergraduate STEM teaching practices: a comparison of instructor self-report instruments

Cody T. Williams[1*], Emily M. Walter[2,3], Charles Henderson[2] and Andrea L. Beach[4]

Abstract

Background: Collecting data on instructional practices is an important step in planning and enacting meaningful initiatives to improve undergraduate science instruction. Self-report survey instruments are one of the most common tools used for collecting data on instructional practices. This paper is an instrument- and item-level analysis of available instructional practice instruments to survey postsecondary instructional practices. We qualitatively analyzed the instruments to document their features and methodologically sorted their items into autonomous categories based on their content. The paper provides a detailed description and evaluation of the instruments, identifies gaps in the literature, and provides suggestions for proper instrument selection, use, and development based on these findings.

Results: The 12 instruments we analyzed use a variety of measurement and development approaches. There are two primary instrument types: those intended for all postsecondary instructors and those intended for instructors in a specific STEM discipline. The instruments intended for all instructors often focus on teaching as well as other aspects of faculty work. The number of teaching practice items and response scales varied widely. Most teaching practice items referred to the format of in-class instruction (54 %), such as group work or problem solving. Another important type of teaching practice items referred to assessment practices (35 %), frequently focusing on specific types of summative assessment items used.

Conclusions: The recent interest in describing teaching practices has led to the development of a diverse set of available self-report instruments. Many instruments lack an audit trail of their development, including rationale for response scales; whole instrument and construct reliability values; and face, construct, and content validity measures. Future researchers should consider building on these existing instruments to address some of their current weaknesses. In addition, there are important aspects of instruction that are not currently described in any of the available instruments. These include laboratory-based instruction, hybrid and online instructional environments, and teaching with elements of universal design.

Keywords: Faculty self-report, Instruments, Postsecondary, Teaching practices

Background

Substantial research has articulated how undergraduate students learn and the instructional practices that best support student learning, including empirically validated instructional strategies (e.g., Chickering & Gamson 1987; Pascarella & Terenzini 1991; 2005). Efforts to transform postsecondary STEM courses to include more of these strategies have had only modest success. One reason for this is that researchers lack shared language and methods for describing teaching practices (Henderson et al. 2011; Beach et al. 2012). As a result, there is a need for documenting tools that describe what teaching practices *actually occur* in college classrooms (American Association for the Advancement of Science [AAAS] 2013).

Surveys are one method to measure the instructional practices of college and university instructors. Self-report surveys can be used alone or in combination with observation to provide a portrait of postsecondary teaching (American Association for the Advancement of Science

* Correspondence: cody.t.williams@wmich.edu
[1]Mallinson Institute for Science Education, Western Michigan University, 1903 W. Michigan Avenue, 49008 Kalamazoo, MI, USA
Full list of author information is available at the end of the article

[AAAS] 2013); these portraits can serve as baseline data for individual instructors, institutions, and faculty developers to plan and enact more effective change initiatives (Turpen & Finkelstein 2009). While self-report surveys are acknowledged as being useful tools for measuring teaching practices, there has been little systematic work characterizing the available instruments.

Ten surveys of postsecondary instructional practices were summarized in a recent report of the American Association for the Advancement of Science (American Association for the Advancement of Science [AAAS] 2013). This report was the result of a 3-day workshop to develop shared language and tools by examining current systematic efforts to improve undergraduate STEM education. Although the report provides an overview of available instruments, it does not examine the design and development of the surveys nor analyze the content and structure of survey items. As a result, it is difficult for researchers to know whether currently available instruments are sufficient or new instruments are needed.

The purpose of this paper is to provide a comparison and content analysis of available postsecondary STEM teaching practice surveys. Our goal is to provide a single resource for researchers to get a sense of the available instruments. We bound our analysis to 10 instruments included in the American Association for the Advancement of Science (AAAS) (2013) report and two instruments that have been released since the report. The AAAS report was developed by a diverse panel of experts in the area of describing college-level STEM. Although we are not aware of any relevant surveys that the AAAS report missed, we are aware of two relevant surveys that have been disseminated since the AAAS report: the Teaching Practices Inventory (TPI; Wieman & Gilbert 2014) and the Postsecondary Instructional Practices Survey (PIPS; Walter et al. 2014). These instruments were included in our analysis because, had they been available at the time, they likely would have been included in the AAAS report (Smith et al. 2014; Walter et al. 2014).

Through our analysis, we seek to characterize the development and administration of the self-report instruments and provide detailed descriptions of their item content (e.g., specific teaching practices) and structure (e.g., clarity, specificity). We also highlight questions that users should consider before adopting or designing an instrument and make suggestions for future work.

Research questions

Our analysis was guided by two research questions:

RQ1. What is the nature of the sample of available surveys that elicit self-report of postsecondary teaching practices?
a. What are the intended populations of the surveys?
b. What measures of reliability and validity were used in the development of the surveys?
c. What is the respondent and administrative burden of the surveys?

RQ2. What teaching practices do the surveys elicit?

Methods

Proper instrument development is essential for a survey to measure correctly its intended subject for its intended demographic (DeLamater et al. 2014). As we considered a comparison of the instruments, we sought to understand the elements essential to their development and administration (RQ1). These elements include the background of the instrument, intended population, respondent and administrative burden, reliability and validity, scoring convention, and reported analyses. These attributes were selected based on commonalities in reported instrument features as well as recommendations in the instrument development literature.

We carefully reviewed the original and related follow-up manuscripts for descriptions of how each instrument employed these features. This section is intended to provide operational definitions for the key features of the instruments; we later describe how these elements were embodied in the instruments we reviewed (see "Results and discussion" section).

Background

Background for an instrument includes details on its original authors, broad development procedure, and a brief description of its content. Where applicable, we include relevant manuscripts associated with the original publication.

Intended population

The intended population of an instrument refers to the group of participants that the instrument was designed to survey (DeLamater et al. 2014).

Respondent and administrative burden

Respondent burden is the amount of time and effort required by participants to complete an instrument. We report estimated time to completion for instruments in their entirety (this may include items other than those related to teaching practice). Administrative burden refers to the demand placed on individuals implementing the instrument. As with respondent burden, the consistency and number of response scales may potentially add to administrative burden.

Reliability and validity

In survey research, it is common to report methods by which reliability and validity were achieved. Reliability is the consistency with which an instrument provides

similar results across items, testing occasions, and raters (Cronbach 1947; Nunnally 1967). There are several commonly reported forms of evidence for instrument reliability, including internal consistency, test-retest, and inter-rater reliability.

Internal consistency addresses whether an instrument is consistent across items and is often reported with Cronbach's alpha (for non-binary surveys). Alpha is a general measurement of the interrelatedness of items, provided there are no covariances, and is dependent on the number of items in the test (Nunnally, 1978).

It is hypothetically possible that the instructional practices on a given survey are not correlated to one another. However, a subset of items on a given survey is typically interrelated in some way. For example, they may have multiple items to get at a particular practice or multiple items designed around a particular construct about teaching (as evidenced by the use of exploratory and confirmatory factor analyses in many studies).

Test-retest reliability refers to the ability of an instrument to produce consistent measurements across testing occasions. Although instructional practices can change over time, some elements could remain consistent.

Inter-rater reliability is the extent to which two or more raters measuring the same phenomenon agree in their ratings. This form of reliability is more common in qualitative work than in survey administration.

Validity is the extent that an instrument measures what it was intended to measure (Haynes et al. 1995). Three commonly reported types of validity are content, construct, and face validity.

Content validity documents how well an instrument represents aspects of the subject of interest (e.g., teaching practices). A panel of subject matter experts is often used to improve content validity through refinement or elimination of items (Anastasi & Urbina 1997). We would expect content validity approaches in all of the surveys we examined.

Construct validity refers to the degree an instrument is consistent with theory (Coons et al. 2000); this is often achieved through confirmatory and/or exploratory factor analyses (Thompson & Daniel 1996). It is not appropriate for every survey to report construct validity since not every survey was developed from a theory base. For example, the TPI (Wieman & Gilbert 2014) was designed as a checklist or rubric of possible teaching practices in a given course. Therefore, as the TPI authors argue, there should be no expectation of underlying constructs.

An instrument has *face validity* if, from the perspective of participants, it appears to have relevance and measures its intended subject. This requires developers to use clear and concise language, avoid jargon, and write items to the education and reading level of the participants (DeLamater et al. 2014). Pilot testing items with a representative sample (e.g., postsecondary instructors) and refining items based on feedback is a common method to improve face validity. We would expect actions to ensure face validity in all of the surveys we examined.

Scoring convention

Scoring convention refers to any procedures used by the instrument authors to score items for the purposes of analyzing participant responses.

Reported analyses

The reported analyses are any statistical procedures used or recommended by the instrument authors to analyze data collected using the instruments. Additionally, the format in which the authors report their data is included here.

Item-level analysis

We undertook a content analysis to understand the aspects of teaching practices measured by each instrument included in the sample (RQ2). Content analysis is a systematic, replicable technique for compressing text (in our case, survey items) into fewer content categories based on explicit coding rules (Berelson 1952; Krippendorff 1980; U.S. General Accounting Office [GAO] 1996; Weber 1990). Content analysis enables researchers to sift through data with ease in a systematic fashion (U.S. General Accounting Office [GAO] 1996) and is a useful technique for describing the focus of individuals or groups (Weber 1990); in our case, we can examine in detail the goals of those surveying the instructional practices of postsecondary instructors. Although content analysis generates quantitative patterns (counts), the technique is methodologically rich and meaningful due to its reliance on explicit coding and categorizing of the data (Stemler 2001).

The analysis began with examining all of the items from the 12 instruments and identifying those related to teaching practices. We ended up with a pool of 320 instructional practice items. Items were excluded from the pool of 320 if they did not capture an instructional practice. We were only interested in analyzing the items that were directly related to instructional practices. The most common type of excluded items were those that elicited only a belief about teaching without the direct implication that the belief informed practice, e.g., "how much do you agree that students learn more effectively from a good lecture than from a good activity?" We did include rationale statements in the analysis as these beliefs directly informed instructional practice, e.g., "I feel it is important to present a lot of facts to students so that they know what they have to learn in this subject."

The first phase of our item-level analysis began with two members of the research team (authors 1 and 2) independently categorizing the 320 items into emergent coarse- and fine-grained codes. The codes were created

based on the content of the items themselves. We designed the codes to be autonomous, that is, one code could not overlap with another. This means that items within coding categories must not only have similar meaning (Weber 1990, p. 37), but codes should be mutually exclusive and exhaustive (U.S. General Accounting Office [GAO] 1996, p. 20). Mutually exclusive categories exist when no item falls between two categories, and each item is represented by only one data point. Generating exhaustive categories is met when the codebook represents all applicable items without exception. For this convention to function, we needed (a) to write code names and code definitions carefully and (b) to sort items into codes based on the single instructional practice best represented by its text.

The second phase of the analysis brought in two additional researchers (authors 3 and 4) to categorize the items using the codebook created by authors 1 and 2. As a four-member research team, we engaged in subsequent rounds of group coding, codebook refinement, and repeated independent coding until an acceptable overall agreement was achieved (82.1 % agreement). The result was 34 autonomous codes in three primary categories: (a) instructional format (20 codes, 138 items), (b) assessment (10 codes, 74 items), and (c) reflective practice (4 codes, 24 items). We define each code and provide a sample item for each in Table 1.

Codebook categories

Instructional format

The instructional format codes refer to items that describe the method by which a course is taught. The codes within the category differ primarily by the primary actors of the instruction, i.e., students versus the instructor. We created three main categories of instructional format codes, including transmission-based instruction, student active, and general practice codes. The transmission-based instruction codes are traditional practices where the instructor is the primary actor. Teaching practices included in this category are lecture, demonstration, and instructor-led question-and-answer. The "student active" codes include a diverse set of practices where students are the primary actors. Example practices in this category are students explaining course concepts, analyzing or manipulating data, completing lab or experimental activities, and having input into course content. The student active codes also included group work practices where two or more students collaborate. The general practice codes consist of practices where there is no designated primary actor, such as connecting course content to scientific research, drawing attention to connections among course concepts, and real-time polling.

Assessment

The assessment codes relate to teaching practices used to determine how well students are learning course content. We created three categories of assessment codes: assignment types, nature of feedback to students, and the nature of assessments. The assignment type codes are various activities assigned to students, i.e., student presentations, writing, and group projects. The "nature of feedback to students" codes refer to how much feedback is given by the instructor to students and the policies enacted by the instructor for how student work is graded. Finally, the "nature of assessment" codes include the types of questions used on summative assessment and the types of outcomes assessed.

Reflective practice

The reflective practice codes are associated with items that ask instructors to think about the big picture of what and how they teach. Additionally, the items ask about how instructors improve their teaching. Example practices include gathering information on student learning to inform future teaching and communicating with students about instructional goals and strategies for success in the course. Also included under the reflective practice codes are items that ask instructors about their rationale behind a particular teaching practice.

Results and discussion

In this section, we review the key features of each instrument. The instruments are described in alphabetical order. Table 2 includes intended population, the number of items and estimated time to completion, and information about reliability and validity for each instrument. Table 3 summarizes the scoring conventions and reported analyses for each instrument. For consistency and ease of explanation, we chose to create a name and acronym for instruments that were not given to them by their original authors. Our titles and acronyms were determined by the STEM discipline of the instrument and original authors' surnames. An asterisk indicates self-generated acronyms.

The "Broad patterns and comparisons" section below includes an overview of the background, intended population, reliability and validity, respondent and administrative burden, scoring convention, and reported analyses across the instruments (RQ1). It then discusses strengths and weakness of the development process used in our sample of instruments. We also consider patterns in the content and structure for the items of each instrument based on our codebook analysis (RQ2). For more in-depth descriptions of each instrument, please see Additional file 1.

Broad patterns and comparisons

What is the nature of the instruments that elicit self-report of postsecondary teaching practices? (RQ1)

Background Almost all of the instruments were developed out of a growing interest to improve undergraduate instruction at a local and/or national scale. Furthermore,

Table 1 Codebook used for content analysis

	Definition	Example item
Instructional format codes		
Connect course content to scientific research	Instructor relates concepts directly to scientific work.	How often did you relate course material to scientific research?
Conceptual framework	Course concepts are related to one another and to outside bodies of knowledge.	Did you use concept maps in your most recent introductory course?
Data	Students analyze or manipulate data.	Students use technology tools to analyze data.
Discussion	Conversation or debate about course concepts among students or between students and the instructor.	In my interactions with students in this subject, I try to develop a conversation with them about the topics we are studying.
Apply or extend	Students extend and apply knowledge to new or challenging situations or to relevant contexts.	What percentage of time on average did your students spend working on problems or projects that required students to seek out new information not previously covered in class?
Explain ideas or concepts	Students' are provided with opportunities to demonstrate their conceptual understanding, process skills, or behaviors.	In your selected course section, how much do you encourage students to explain course material to other students?
Vague instructional format	Forms of instructional practice described broadly or vaguely and not described by another code.	Activities are used to help students learn statistics.
Generic group work	Group work described broadly and without specific product or context.	How often do you put students into pairs or small groups for most of a class period to answer questions or solve problems?
Group problem solving	Group work involving solving problems.	Students engaged in structured collaborations to solve problems.
Group assessment	Students complete in-class tests or quizzes in groups or complete out-of-class assignments in groups with the goal of creating a formal product.	Did you use group and team projects that produce a joint product?
Lab or experiment	Students are doing lab or experiment-based tasks.	How frequently did you use fieldwork in your most recent introductory course?
Problem solving	Students solve problems individually in class. Problems could be conceptual or computational.	How frequently did students solve/discuss quantitative/ mathematical problems in the lecture portion of course?
Process of science	Students engage in a scientific investigation with varying degrees of guidance from the instructor.	How frequently did students design experiments/activities in the lecture portion of course?
Read the primary literature	Students read primary literature that may or may not be provided by the instructor.	In your interactions with undergraduates, how often do you encourage them to look up scientific research articles and resources?
Reflection	Strategies to encourage students to think about what they know and why they know it.	In how many of the courses that you teach do you use reflective writing/journaling?
Students have input	Students have some level of control over what they learn and/or the order in which they learn.	Students posed and solved their own problem.
Lecture	The instructor speaks and students passively listen.	How often did you use extensive lecturing (more than 15 min per session without breaks for questions or active engagement of students)?
Demonstration/ example	The instructor displays a phenomenon or how to do a task.	How frequently did you use lecture with demonstration in your most recent introductory course?
Question and answer (Q&A)	The instructor posing questions to the class as a whole and receiving individual student responses.	How often do you address questions to the class as a whole?
Real-world context	Students work within a real-world context OR the instructor relates content to real-world examples.	Did you use service learning, co-op experiences, or assignments requiring interactions with the community or business/industry?
Assessment practice codes		
Assessment of specific outcomes	Assessment designed to evaluate students' understanding of course concepts or ability to perform a given task.	One use of my assessments is to reveal whether students are using statistical language properly.
Formative assessment	Practices that elicit, build upon, or evaluate students' knowledge and ideas prior to a summative assessment.	Have you used a pre-test of prior knowledge?
Vague assessment practice	Forms of assessment described broadly or vaguely and not described by another assessment code.	Did you use exams in your most recent introductory course?

Table 1 Codebook used for content analysis *(Continued)*

Grading policy	Instructors' set of rules about how students are assessed and how those policies are communicated to students.	In how many of the courses that you teach do you use grading on a curve?
Group assessment	Students complete in-class tests or quizzes in groups.	My assessments include a component in which students are required to collaborate (e.g., group project, group quiz)
Nature of feedback to students	Details of how (e.g., frequency, immediacy, and level of detail) instructor provides feedback on student work.	In how many of the courses that you teach do you use electronic quizzes with immediate feedback in class?
Nature of questions (type and variety)	Type or format of question and/or the use of a variety of question types on summative assessments.	Have you used essay questions?
Peer evaluation	Students provide feedback on each other's work.	Have you used student evaluations of each other's work?
Student presentations	Students present their ideas or findings to the class in a formal presentation.	Did you use oral presentations in your most recent introductory course?
Writing	Students produce a formal writing product.	How often did you use writing assignments (reflective writing, journals, essays, reports)?
Reflective practice codes		
Reflective practice	Instructors gather and use information to inform their future instruction.	Have you used mid-semester course evaluations regarding your teaching?
Instructional goals	Learning goals are provided to students.	In your undergraduate courses, to what extent do you clearly explain course goals and requirements?
Strategies for success	Instructor recommendations that encourage behavior that will lead to student success in the course.	In your interactions with undergraduates, how often do you encourage them to ask questions in class?
Orientation toward teaching	The instructors' rationale behind particular instructional, curriculum, or assessment strategies.	I feel it is important to present a lot of facts to students so that they know what they have to learn for this subject.

eight of the 12 surveys we reviewed have been published or revised since 2012, heralding a movement among the research community to measure the state of undergraduate education.

Intended population Four of the instruments we reviewed span all postsecondary disciplines (Faculty Survey of Student Engagement (FSSE), Higher Education Research Institute (HERI), National Study of Postsecondary Faculty (NSOPF), PIPS). The remaining instruments are designed for STEM faculty, including physics (Henderson & Dancy Physics Faculty Survey (HDPFS)) and engineering faculty (Borrego Engineering Faculty Survey (BEFS), BREFS), chemistry and biology (Survey of Teaching Beliefs and Practices (STEP)), geosciences (On the Cutting Edge Survey (OCES)), statistics (Statistics Teaching Inventory (STI)), and science and mathematics (TPI). There are no instruments designed specifically for technology postsecondary instructors, with the exception of an instrument to measure integration of technology into postsecondary math classrooms (Lavicza, 2010). However, this instrument focuses on use of particular technologies and not particular teaching practices.

Administrative and respondent burden There is great variability in the number of items on the surveys we reviewed (84.4 ± 72.7). Lengthy surveys, such as the FSSE (130 items), HERI (284 items), NSOPF (83 items), TPI (72 items), and STEP (67 items), may cause participants to develop test fatigue, i.e., become bored or not pay attention to how they respond (Royce 2007).

The number of teaching practice items (26.7 ± 14.2) and proportion of teaching practices in the overall instrument (43.4 ± 26.1 %) also vary widely. This may be problematic for administrators seeking only to elicit teaching practices of respondents. Furthermore, although teaching practice items could be pulled out from a larger survey, this can impact the construct validity of the instrument.

The instruments with the lowest proportion of teaching practice items are national interdisciplinary surveys designed to assess multiple elements of the faculty work experience: FSSE (17.7 % instructional practice items), HERI (12.3 %), and NSOPF (12.0 %). In contrast, the remaining (mostly discipline-specific with the exception of PIPS) instruments focus more items on instructional practices: TPI (83.3 %), PIPS (72.7 %), HDPFS (65.6 %), OCES (63.0 %), Approaches to Teaching Inventory (ATI) (56.3 %), STI (42.0 %), and STEP (34.9 %). The exception to this pattern is Southeastern University and College Coalition for Engineering Education (SUCCEED), with only 17.9 % of its items devoted to instructional practices.

There are also a variety of scales employed by the instruments we analyzed (Table 4). Many used a 5-point response scale (e.g., BEFS, PIPS, STI, SUCCEED, TPI), but others use 3-point (STEP, NSOPF, SUCCEED, TPI), 8-point (FSSE), and binary scales (OCES, STI, SUCCEED, TPI). Response scales are an important consideration in instrument development, as is an explicit rationale for given scales in development documents. Five-point scales

Table 2 Instrument key features (part 1)

Instrument	Intended population	Number of items and time to completion	Reliability and validity
Approaches to Teaching Inventory (ATI) (Trigwell & Prosser 2004)[a]	Postsecondary faculty	9 teaching practice items 16 total items No time to complete given	• Two pilots • Categories of teaching approaches based on interviews with 24 science teachers • Principal component analysis with varimax rotation (16 items, 2 sub-scales) • Confirmatory factor analysis
Borrego Engineering Faculty Survey (BEFS) (Borrego et al. 2013)	Engineering faculty members teaching engineering science courses	27 teaching practice items No total number of items given No time to complete given	• Based on previously developed survey • Research-based instructional strategies (RBIS) and critical components of research-based instructional strategies based on expert literature review • Reliability $\alpha = 0.921$ (internal consistency)
Faculty Survey of Student Engagement (CPRIU 2012)	Postsecondary faculty in the US	23 teaching practice items 130 total items 15–20 min	• 2013 version refined based on user feedback • Exploratory and confirmatory factor analysis (9 constructs)
Henderson & Dancy Physics Faculty Survey (HDPFS) (Henderson & Dancy 2009)	Physics faculty teaching introductory quantitative physics courses	40 teaching practice items 61 total items No time to complete given	• RBIS and critical components of research-based instructional strategies based on expert literature review • Developed in consultation with experts at the American Institute of Physics Statistical Research Center
Higher Education Research Institute Faculty Survey (HERI) (Hurtado et al. 2012)	Postsecondary faculty in the US	35 teaching practice items 284 total items 25 min	• Item pools created based on literature • Exploratory factor analysis (11 constructs)
National Study of Postsecondary Faculty (NCES 2004)	Postsecondary faculty in the US	10 teaching practice items 83 total items 30 min	• Based on the 1999 version • Revised based on comments from project Technical Review Panel, previous respondents, government officials, and postsecondary researchers • Revised instrument was then field tested
On the Cutting Edge Survey (OCES) (Macdonald et al. 2005)	Geosciences faculty teaching undergraduate geoscience courses	29 teaching practice items 46 total items No time to complete given	• Based on a previous biology faculty instrument (First II) and Using Data in the Classroom (Manduca & Mogk, 2002) • Piloted with 16 faculty at an American Geophysical Union meeting
Postsecondary Instructional Practices Survey (PIPS) (Walter et al. 2014)	All postsecondary instructors	24 teaching practice items 33 total items 10 min	• Initial items based on the literature • Reviewed by 4 educational researchers • Pilot tested with 827 instructors • Exploratory and confirmatory factor analysis (2 and 5 factor solutions)
STEP Survey (Marbach-Ad et al. 2012)	Chemistry and biology faculty, graduate teaching assistants, and undergraduates	30 teaching practice items 86 total items No time to complete given	• Initial items based on the literature • Pilot testing and review by experts from the sciences, education, and psychology
Statistics Teaching Inventory (STI) (Zieffler et al. 2012)	Statistics faculty across disciplines and institutions	21 teaching practice items 50 total items No time to complete given	• Initial items based on the curriculum Guidelines for Assessment and Instruction in Statistics Education • Piloted with members of the statistics education community • Revised instrument was used for think aloud interviews (reliability) • Another small-scale pilot was done with local statistics educators
SUCCEED Survey (Brawner et al. 2002)	Engineering faculty at eightUS universities	12 teaching practice items 67 total items No time to complete given	• Initial items based on the literature • Piloted by project leadership team
Teaching Practices Inventory (TPI) (Wieman & Gilbert 2014)	Postsecondary science and math instructors	60 teaching practice items 72 total items 13 min	• Initial items developed based on authors' experience and previous teaching practice rubric • Two pilot tests with 150 and 179 instructors • Multiple rounds of expert review • Scoring procedure based on the literature

[a]We recognize that a revised version of the ATI was released in 2005 (Trigwell, Prosser, & Ginns 2005). This version of the ATI was intended to further improve its utility for promoting discussion among teachers about teaching approaches. We decided to focus our review on the 2004 ATI since the 2005 version was modified in focus and because the 2004 version is more cited than its 2005 counterpart (285 to 89 citations, respectively)

Table 3 Instrument key features (part 2)

Instrument	Scoring convention	Reported analyses
ATI	Each item scored from 1 to 5 from lower to higher frequency	Correlational studies of ATI scores and teaching and learning factors
BEFS	Distinguished between users and non-users of research-based instructional strategies (RBIS) and their critical components	Response trends in percentagesCompared use of RBIS and critical components with Fisher's exact test and chi-square
FSSE	None reported	Frequency distributions of responses for each item
HDPFS	None reported	Frequency distributions of responses for each item
HERI	Construct scores estimated based on pattern of responses for construct	Frequency distributions of responses for each itemConstruct scores reported in low, average, and high categories
NSOPF	None reported	Frequency distributions of responses for each item
OCES	None reported	Frequency distributions of responses for each item
PIPS	Items scored 0–4, high scores for greater descriptiveness of one's teachingConstruct scores = total points for construct items/total possible points for construct multiplied by 100	ANOVA and independent *t* tests used to compare demographic groupsCorrelational analysis used to explore relationships between PIPS scores and aspects of teaching
STEP	Instructional format and perceptions of teaching items scored from 1 to 5, higher scores assigned to greater importance or use	Frequency distributions of responses for each item*t* tests used to compare means for individual items
STI	Response choices assigned score 0–1, higher scores for more reformed practices	Frequency distributions of responses and mean responses for each item
SUCCEED	Teaching practice items scored 0–4, higher scores for more frequent use	One-way ANOVA to compare mean scores
TPI	Scores 1–3 assigned by item for yes responses. Higher scores for literature supported reformed practices	Whole instrument mean scores, standard deviation, and course enrollment weighted mean scores

are generally recommended to maximize variance in responses, unless there is a compelling reason not to use such a scale (Bass et al. 1974; Clark & Watson 1995). Despite recommendations in the literature, authors rarely voiced their rationale for scale choice. Notable exceptions to this are the STI (Zieffler et al. 2012) and PIPS (Walter et al. 2014), which document rationale behind selecting a scale.

Scoring convention Seven of the instruments reported some form of scoring system. In general, scoring is done on a positive scale with higher scores given to responses indicating greater importance or use of reformed teaching practices. Providing scoring systems for an instrument can help users make sense of large data sets and produce more consistent data sets across implementations.

Reported analyses The majority of the instruments reported descriptive statistics such as frequency distributions, means, and standard deviations. A few instruments (BEFS, PIPS, STEP, SUCCEED) reported mean comparisons using common statistical test such as independent *t* tests, ANOVA, and chi-square. Some instruments (ATI, PIPS) also reported correlational analysis between instrument scores and various aspects related to teaching and learning.

Areas for improvement and strengths related to the development of existing instrumentation

Face validity It is key that an instrument makes sense and appears to measure its intended concept from the perspective of the participant (DeLamater et al. 2014). This requires avoiding jargon-based (e.g., inquiry, problem solving), overly complex, and vague statements. Although 8 of the 12 instruments were pilot tested and revised before wide implementation, we coded vague teaching practice items in all instruments except the ATI, regardless of whether they were pilot tested (see Additional file 2). "Vague" items by our definition could not be described by another instructional format or assessment code, because they were too broadly described. For example, "How often did you use multimedia (e.g., video clips, animations, sound clips)?" (Marbach-Ad et al. 2012). Similarly, many instruments included double-barreled (or multi-barreled) items, which described two or more concepts in a single question. For example, "In your selected course section, how much does the coursework emphasize applying facts, theories, or methods to practical problems or new situations?" (Center for Postsecondary Research at Indiana University [CPRIU]). These items can be problematic for participants to answer and can provide data that is difficult to interpret for researchers (Clark & Watson 1995). We encourage users to look for and identify vague items in any instrument, as these items may reduce face validity and fail to produce meaningful data.

Content validity Seven of the instruments we reviewed have documented use of an outside panel of experts to improve content validity (BEFS, HDPFS, OCEA, PIPS, STEP, STI, and TPI). In particular, we highlight the efforts of the authors of the STI (Zieffler

Table 4 Nature of the scales used by the instruments

Scale type	Sample item	Instrument
Binary scales		
Agree, disagree	How often do you assign at least one major team project?	SUCCEED
3-point scales		
Do not use; use and counts towards student grade; use and does not count towards student grade	Have you used pre-tests of prior knowledge?	STEP
Used in all classes, used in some classes, not used	Did you use a multiple-choice midterm or final exam?	NSOPF
In every course, I teach; in some but not all courses, I teach; never	How often do you assign at least one major team project?	SUCCEED
0–10 %, 10–25 %, more than 25 %	Considering the time spent on the major topics, approximately what fraction was spent on *process* by which the theory/model/concept was developed?	TPI
4-point scales		
Very important; important; somewhat important; not important	To what extent do you clearly explain course goals and requirements?	FSSE
All, most, some, none	In how many of the courses that you teach do you use multiple drafts of written work?	HERI
Never used on tests; used occasionally on tests; used frequently on tests; used on all tests	On tests and quizzes, how often did you use conceptual questions?	HDPFS
Always; usually; sometimes; never	How often do you give students study guides before tests?	SUCCEED
5-point scales		
I currently use all or part of it; I have used all or part of it in the past; I am familiar with it but have never used it; I have heard the name but do not know much else about it; I have never heard of it	Please indicate your level of familiarity with inquiry learning.	BEFS, HDPFS
0 %, 1–25 %, 26–50 %, 51–75 %, and 76–100 %	Please indicate what percentage of time on average your students spent/spend working on problem sets or projects in pairs or small groups.	BEFS
0–20 %, 20–40 %, 40–60 %, 60–80 %, 80–100 %	Fraction of typical class period you spend lecturing.	TPI
Never; seldom; some of the time; most of the time; all of the time	Please rate the extent that real data sets are used during instruction.	STI
Not used; once per semester; a few times a semester; most class sessions; almost every class session	How often did you use class discussions?	STEP
Never; once or twice; several times; weekly; for nearly every class	Please indicate how frequently you used lecture with demonstration.	OCES
Not at all descriptive of my teaching; minimally descriptive of my teaching; somewhat descriptive of my teaching; mostly descriptive of my teaching; very descriptive of my teaching	I require students to work together in small groups.	PIPS
Only rarely; sometimes; about half the time; frequently; always	I structure this subject to help students pass the formal assessment items	ATI
Never; one or more times a semester; one or more times a month; one or more times a week; every class	How often do you address questions to the class as a whole?	SUCCEED
6-point scales		
Never; Once or twice; Several Times; Weekly; For Nearly Every Class; Multiple Times Every Class	How frequently did you use traditional lecture in the lecture portion of the course?	HDPFS
8-point scales		
0 %, 1–9 %, 10–19 %, 20–29 %, 30–39 %, 40–49 %, 50–74 %, 75 % or more	In your selected course section, about what percent of class time is spent on lecture?	FSSE

et al. 2012), for their iterative review process utilizing statistics education community members and NSF project advisors.

Construct validity Construct validity is the least addressed component of validity in the instruments we reviewed. Only the ATI (2 constructs), FSSE (9 constructs),

HERI (11 constructs), and PIPS (2 or 5 constructs) have documented analyses of how items grouped together in factor or principal components analyses. Furthermore, only the ATI, FSSE, and PIPS use confirmatory factor analyses to sort items into a priori categorizations. To this end, we add that none of the instruments build upon a specific educational theory nor generate a theoretical framework for the nature of postsecondary instructional practice.

Reliability Only two of the available instruments (BEFS and FSSE) cite reliability values by construct. All other instruments fail to provide reliability statistics, bringing into question the precision of their results. Furthermore, none of the instruments we reviewed provided test-retest reliability statistics. We encourage future users of the instruments to consider longitudinal studies that would allow for the publication of these values.

Development process We were surprised by the lack of documentation available for the development process of the instruments we reviewed. How items were generated, revised, and ultimately finalized was often not apparent. Survey development should be a transparent process, available online if not in manuscript. The ATI and STI are good examples of detailed methodological processes, providing extensive detail from development of the initial item pool, item refinement, and pilot testing to data analyses and ongoing revisions. Rationale should also be provided for item scales, with the goal of avoiding unjustified changes in scale among item blocks. We recommend referencing the psychometric literature (e.g., Bass et al. 1974; Clark & Watson 1995) to provide support for the use of particular scales.

What teaching practices do the instruments elicit? (RQ2)

As we examined all of the instruments in our sample, the majority had the largest number of their items focused on instructional format (BEFS, HDPFS, OCES) or a combination of instructional format and assessment (FSSE, PIPS, STI, STEP, SUCCEED, TPI). Other instruments had a variety of different foci. The ATI has a nearly equal number of reflective practice items ($n = 4$) to instructional format items ($n = 5$), and the NSOPF devotes almost all of its 10 teaching practice items to assessment practice ($n = 9$). Only the HERI has equal proportions of instructional format, assessment, and reflective practice items, although these items are a subset of 284 total questions on the instrument. Figure 1 provides a breakdown of item types by instrument.

Across the full 320-item pool, most items were coded into the instructional format category (see Additional file 2 for a full tabulation of codes). These 174 items most often referred to discussions ($n = 17$), group work ($n = 16$), students doing problem solving activities ($n = 16$), instructor demonstration/example ($n = 11$), real-world contexts ($n = 12$), real-time polling ($n = 9$), and using quantitative approaches to manipulate or analyze data ($n = 9$). Rarely did items describe instruction in a lab or field setting ($n = 6$). In addition, the lab-specific items did not reflect current reforms in laboratory instruction (e.g., avoiding verification-based activities or allowing flexibility in methods; Lunetta et al. 2007).

Assessment practice items ($n = 111$) focused primarily on the nature of summative assessments. Items usually referred to instructor grading policy ($n = 20$), the format of questions on summative assessments (e.g., multiple-choice, open-ended questions) ($n = 19$), formative assessment ($n = 12$), or the general format of summative assessments (e.g., midterms, quizzes) ($n = 11$). The remaining assessment items primarily referred to student term papers ($n = 10$), group assessments ($n = 7$), student presentations ($n = 7$), content assessed on summative assessments ($n = 6$), the nature of feedback given to students ($n = 6$), and peer evaluation of assessments ($n = 4$). There

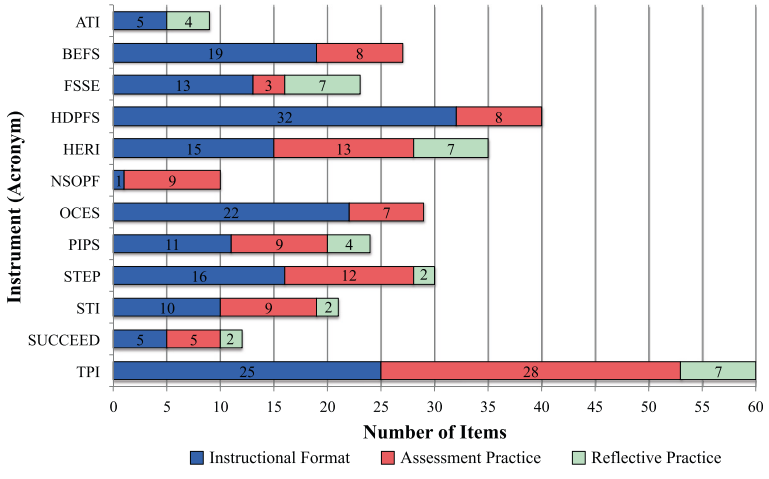

Fig. 1 Instrument items per coding category. Number of items per code category for postsecondary instructional practice surveys

is a lack of instruments that explicitly refer to formative assessment practices, those that elicit, build upon, or evaluate students' prior knowledge and ideas (Angelo & Cross 1993). While there were 12 total items referring to formative assessment, over half of the formative assessment items came from one instrument (TPI). Although the nine items sorted into the "real-time polling" code could refer to formative assessment, the use of clickers and whole class voting does not imply formative use.

We also looked specifically at the discipline-based instruments in our sample including the BEFS (engineering), HDPFS (physics), OCES (geosciences), STEP (chemistry and biology), STI (statistics), and SUCCEED (engineering). Most of the discipline-based instruments focused the majority of their items on instructional format. The SUCCEED and the STI are exceptions in that they are evenly split between instructional format and assessment. The instructional format items across the discipline-based instruments most commonly focused on group work ($n = 14$), students analyzing data ($n = 9$), discussion ($n = 6$), and lecture ($n = 5$). Some of the instruments dedicated a substantial amount of their instructional format items to particular practices. For example, the HDPFS ($n = 7$) and OCES ($n = 5$) both have several items related to problem solving. The OCES ($n = 5$) and STI ($n = 3$) have items focused on having students quantitatively analyze datasets. In addition, BEFS has a particular focus on providing a real-world context for students ($n = 4$) and group work ($n = 4$). The HDPFS is also noteworthy for being the only discipline-based instrument with multiple items ($n = 3$) related to laboratory teaching practices. Only one other discipline-based instrument, the OCES, has a single item related to the laboratory.

The discipline-based instruments also had a secondary focus on assessment practices. The most common assessment items across the instruments were those related to the nature of the questions included on course assessments. In particular, the HDPFS authors dedicated the majority of their assessment items ($n = 6$) to the nature of assessment questions. This being said, there were two instruments that had a unique focus for their assessment practice items. The STI has six items (out of nine) related to including specific content on assessments, while the SUCCEED has three items (of five) focused on group assessments.

None of the disciplinary instruments had many reflective practice items. Three instruments had no reflective practice items. Two minor exceptions are the STEP and the SUCCEED, which both had two items aimed at whether learning goals are provided to students.

Conclusions

Although many of the instruments have development and/or psychometric issues, no instrument is wholly problematic. To conclude the paper, we return to our research questions and provide recommendations for users and developers of postsecondary teaching practice surveys.

What is the nature of the instruments that elicit self-report of postsecondary teaching practices? (RQ1)

The majority of instruments we reviewed were designed for particular STEM disciplines. Outside of large national instruments, there are few instruments designed for measuring teaching practices across disciplines. In addition, there is considerable variability in overall instrument length, the proportion of teaching practice items, and response scales. All of these aspects should be taken into account to maximize participants' ease of completing the instrument and researchers' interpretations of the data produced.

Considerations for users and developers

The purpose of this paper has been to analyze and compare available instruments, in part so that readers have a sense of direction when determining how to measure instructional practices in their given context. Based on this experience, we are able to identify questions for potential users and developers of postsecondary instructional practice instruments. This is not a set of research questions but rather questions to consider prior to implementation. For more specific recommendations for quality test administration, consider the guidelines published by the International Test Commission (International Test Commission [ITC] (2001)).

Consideration 1: is there an established instrument?

We consider the first step to finding or developing a postsecondary teaching practice instrument to be an examination of what is currently available. We have created a flowchart (Fig. 2) to help users distinguish among the basic features of available instruments. Please note that this chart is a first step to navigating the sea of available instruments. It should *not* be interpreted as a recommendation for any of the instruments without deeper examination of the validity, reliability, content, and clarity of an instrument.

Consideration 2: is the instrument valid and reliable?

Upon confirmation that an instrument is appropriate for a particular audience, context, and research questions, the instrument should be assessed to determine if it measures what it was intended to target (validity; Haynes et al. 1995) and produces repeatable and precise results (reliability; Cronbach 1947; Nunnally 1967). We report common methods to achieve validity and reliability earlier in the manuscript (see Key Features of the Instruments), and we summarize the methods used for each instrument in Table 2. If validity and reliability have been accounted for, a user can have some confidence in the results produced by an instrument. Keep in mind that not all measures of

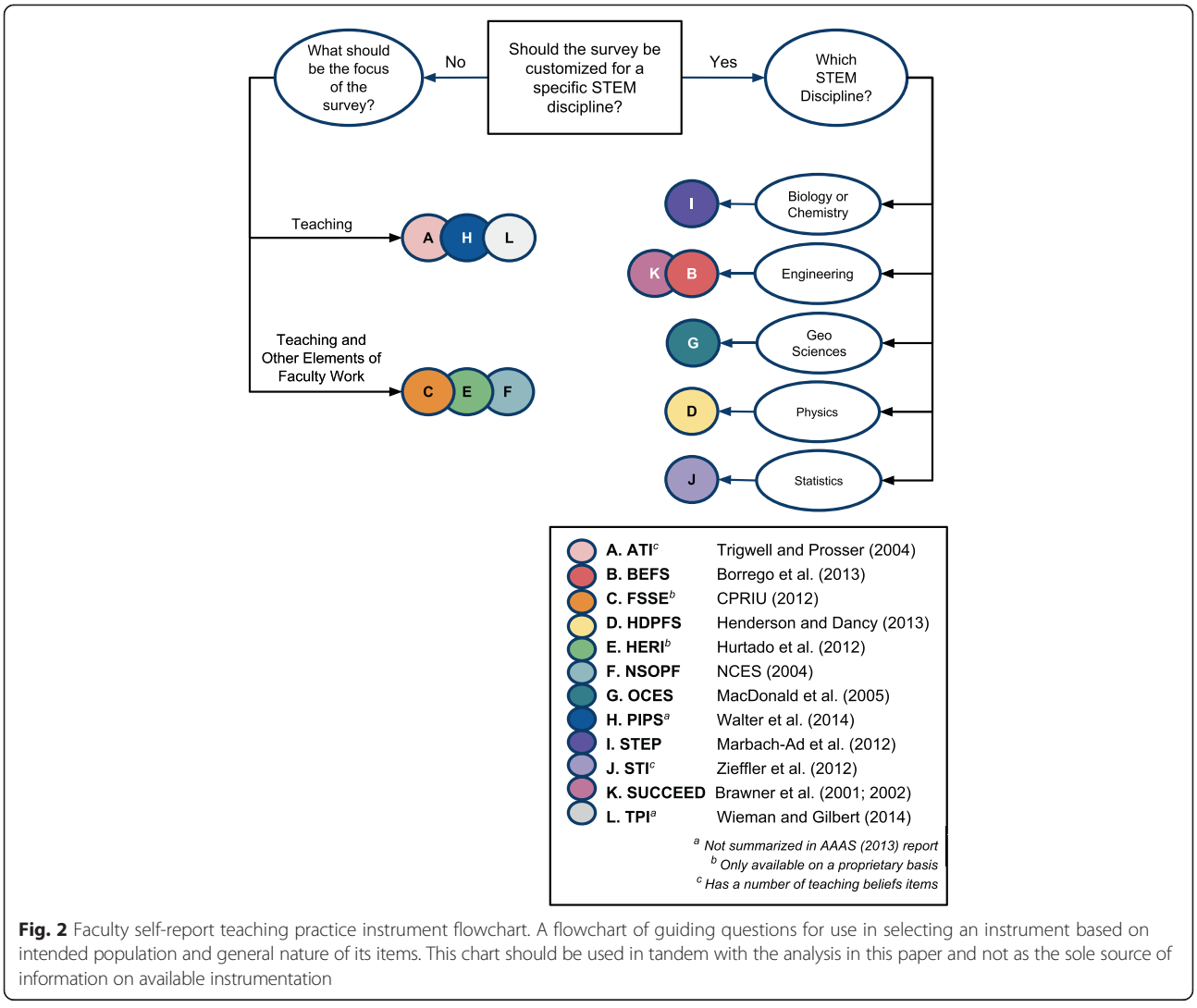

Fig. 2 Faculty self-report teaching practice instrument flowchart. A flowchart of guiding questions for use in selecting an instrument based on intended population and general nature of its items. This chart should be used in tandem with the analysis in this paper and not as the sole source of information on available instrumentation

validity and reliability are appropriate, depending on the goals of the instrument and how it was developed.

Consideration 3: what response scale(s) does the instrument use?

Inconsistent and unjustified item scales may add to administrative burden of a test and may contribute to test fatigue (Royce 2007). We recommend careful examination of item scales including number of response options (see Bass et al. 1974) and use of a neutral point on the scale. Forcing agreement or disagreement through eliminating a neutral option may avoid an increase in participants claiming "no opinion" when they actually have one (Bishop 1987; Johns 2005).

Consideration 4: will you modify or adapt the instrument?

Should a user decide that an instrument is valid, reliable, and acceptable for their intended audience, we recommend that the survey be administered in its entirety and without modifying the items. Gathering data in this controlled way enables the comparison of data with others that have used the instrument and preserves construct validity (van de Vijver 2001). Deviations from these conditions should be reported as constraints on the interpretation of results. We note that using a complete instrument may be more challenging for users interested in the FSSE, HERI, NSOPF, and/or SUCCEED, as these surveys have a large number of non-teaching practice items.

Consideration 5: do you plan to develop a new instrument?

Should the current instrumentation be insufficient for your needs, we recommend that instruments are created in the most methodological and transferable way possible (e.g., Rea & Parker 2014). Keep and disseminate detailed records of your development process, testing, and analyses. Communicate with other research groups for compatibility, comparability, and further reliability and validity testing. Since there has been little work to compare data gathered from the same population using different teaching practice

instruments, we suggest gathering data using both the new instrument and a reliable and valid existing instrument to see how the instruments elicit teaching practices in similar or unique fashions.

What teaching practices do the instruments elicit? (RQ2)

The bulk of the teaching practice items across the instruments reviewed were focused on instructional format and or assessment practice. Two important areas that seem to be missing from many of the instruments are lab instructional practices and formative assessment. These are both areas that should be addressed in future instrument development.

Recommendations for future research

As discussed in this paper, many instruments currently exist for describing postsecondary teaching practices. More work is certainly needed to further refine these instruments and other similar instruments. More importantly, though, the field currently lacks instrumentation for measuring teaching practices in laboratory and online settings.

Measuring instructional practices in online courses

Despite widespread and increasing adoption of online learning approaches (Johnson et al. 2013), there are no comprehensive surveys of online teaching practices nor an objective set of descriptors to classify online teaching practices. This is not to say we do not know what makes effective online instruction. Significant effort by instructional designers, faculty developers, and online platform providers has generated checklists and rubrics of best practices (e.g., Quality Matters, BlackBoard Exemplary Course Program Rubric, MERLOT Evaluation Standards for Learning Materials).

However, best practice rubrics are designed for self-reflection or peer evaluation. They are not designed to consistently and precisely measure the same instructional practices over separate administrations, nor are they confirmed to measure what they intend. For proper comparisons among data sets and accurate results, valid and reliable instruments should be designed to measure instructional practices in online settings.

Laboratory instructional practices

Like online course settings, we find the surveys available for face-to-face classrooms to be missing an element that describes components of effective laboratory teaching. This includes avoiding verification-based activities and allowing flexibility in methods (e.g., Lunetta et al. 2007).

Inclusivity

Lastly, we see little discussion of teaching strategies specific to improving outcomes for many groups of students that are typically underrepresented in STEM disciplines, such as students with disabilities or underprepared students. Such students make up an increasing proportion of the college student population. We consider many reform-based instructional strategies to include components of universal design (Scott et al. 2003); universal design requires an intentional approach to a variety of human needs and diversity. Some universal design elements may be elicited through items on existing instruments, including items that highlight a community of learners, flexibility in teaching methods, and tolerance for student error on assessments. Other elements, including the intentionality to use methods that address the needs of diverse learners, are not as apparent in the current instrumentation. We encourage developers to consider elements of universal design when generating survey items.

Additional files

Additional file 1: Individual instrument summaries. In this appendix, we review the key features of each instrument. The instruments are described in alphabetical order. The review includes the background, intended population, reliability and validity, respondent and administrative burden, scoring convention, and reported analyses for each instrument. (PDF 209 kb)

Additional file 2: Coding data. In this appendix, we include frequency counts for each of our codes. Counts are given for each code for each instrument. There are also totals provided across all of the instruments for each code. (XLSX 13 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CW was primarily responsible for writing the manuscript. CW and EW conducted the majority of the instrument analysis. All authors were involved in coding the instrument items. EW, CH and AB provided feedback throughout. All authors read and approved the manuscript.

Acknowledgements

This paper is based upon work supported by the National Science Foundation under Grant No.1256505. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the NSF.

Author details

[1]Mallinson Institute for Science Education, Western Michigan University, 1903 W. Michigan Avenue, 49008 Kalamazoo, MI, USA. [2]Physics Department, Western Michigan University, 1903 W. Michigan Avenue, 49008 Kalamazoo, MI, USA. [3]Department of Biology, California State University, 2555 E. San Ramon Ave, 93740 Fresno, CA, USA. [4]Department of Educational Leadership, Research and Technology, Western Michigan University, 1903 WMichigan Ave, 49008 Kalamazoo, MI, USA.

Received: 15 January 2015 Accepted: 15 October 2015
Published online: 29 October 2015

References

American Association for the Advancement of Science [AAAS]. (2013). *Describing and measuring undergraduate STEM teaching practices*. Washington, DC: Author.

Anastasi, A, & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River: Prentice Hall.

Angelo, TA, & Cross, KP. (1993). *Classroom assessment techniques: a handbook for college teachers* (2nd ed.). San Francisco: Jossey-Bass.

Bass, B, Cascio, W, & O'Connor, E. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59*, 313.
Beach, AL, Henderson, C, & Finkelstein, N. (2012). Facilitating change in undergraduate STEM education. *Change: The Magazine of Higher Learning, 44*(6), 52–59. doi:10.1080/00091383.2012.728955.
Berelson, B. (1952). *Content analysis in communication research*. Glencoe: Free Press.
Bishop, GF. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly, 51*, 220–232.
Borrego, M., Cutler, S., Prince, M., Henderson, C., & Froyd, J. (2013). Fidelity of implementation of Research-Based Instructional Strategies (RBIS) in engineering science courses. *Journal of Engineering Education, 102*(3). doi:10.1002/jee.20020
Brawner, CE, Felder, RM, Allen, R, & Brent, R. (2002). A survey of faculty teaching practices and involvement in faculty development activities. *Journal of Engineering Education – Washington, 91*, 393–396.
Center for Post-secondary Research at Indiana University [CPRIU]. (2012). Faculty Survey of Student Engagement (FSSE). Retrieved from: http://fsse.indiana.edu/pdf/2012/FSSE12_TS.pdf.
Chickering, AW, & Gamson, ZF. (1987). *Applying the seven principles for good practice in undergraduate education*. San Francisco: Jossey-Bass.
Clark, L, & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment, 7*, 309.
Coons, SJ, Rao, S, Keininger, DL, & Hays, RD. (2000). A comparative review of generic quality-of-life instruments. *PharmacoEconomics, 17*(1), 13–35.
Cronbach, LJ. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12*, 1–16.
DeLamater, JD, Myers, DJ, & Collett, JL. (2014). *Social psychology* (8th ed.). Boulder: Westview Press.
Haynes, SN, Richard, DCS, & Kubany, ES. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment, 7*, 238–247. doi:10.1037/1040-3590.7.3.238.
Henderson, C, Beach, AL, & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: an analytic review of the literature. *Journal of Research in Science Teaching, 48*, 952–984. doi:10.1002/tea.20439.
Henderson, C., & Dancy, M. (2009). The impact of physics education research on the teaching of introductory quantitative physics in the United States. *Physical Review Special Topics: Physics Education Research, 5*(2). doi: 10.1103/PhysRevSTPER.5.020107
Hurtado, S., Eagan, K., Pryor, J. H., Whang, H., & Tran, S. (2012). *Undergraduate teaching faculty: the 2010–2011 HERI faculty survey*.
International Test Commission [ITC]. (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93–114. Retrieved from: http://www.tandfonline.com/doi/abs/10.1207/S15327574IJT0102_1.
Johns, R. (2005). One size doesn't fit all: selecting response scales for attitude items. *Journal of Elections, Public Opinion, & Parties, 15*, 237–264. doi:10.1080/13689880500178849.
Johnson, L, Adams Becker, S, Estrada, V, & Martín, S. (2013). *Technology outlook for STEM+ education 2013–2018: an NMC horizon project sector analysis*. Austin: The New Media Consortium.
Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Newbury Park: Sage.
Lavicza, Z. (2010). Integrating technology into mathematics teaching at the university level. *ZDM Mathematics Education, 42*, 105–119. doi:10.1007/s11858-009-0225-1.
Lunetta, VN, Hofstein, A, & Clough, MP. (2007). Learning and teaching in the school science laboratory: an analysis of research, theory, and practice. In SK Abell & NG Lederman (Eds.), *Handbook of research on science education* (pp. 393–441). Mahwah: Lawrence Erlbaum.
MacDonald, RH, Manduca, CA, Mogk, DW, & Tewksbury, BJ. (2005). Teaching methods in undergraduate geoscience courses: results of the 2004 On the Cutting Edge Survey of U.S. faculty. *Journal of Geoscience Education, 53*, 237–252.
Manduca, C. A., & Mogk, D. W. (2003). Using data in undergraduate science classrooms. Northfield, MN. Retrieved from http://d32ogoqmya1dw8.cloudfront.net/files/usingdata/UsingData.pdf.
Marbach-Ad, G, Schaefer-Zimmer, KL, Orgler, M, Benson, S, & Thompson, KV. (2012). *Surveying research university faculty, graduate students and undergraduates: skills and practices important for science majors*. Vancouver: Paper presented at the annual meeting of the American Educational Research Association (AERA).
National Center for Education Statistics [NCES]. (2004). *National Study of Postsecondary Faculty (NSOPF)*. National Center for Education Statistics. http://nces.ed.gov/surveys/nsopf/pdf/2004_Faculty_Questionnaire.pdf
Nunnally, JC. (1967). *Psychometric theory*. New York: McGraw-Hill.
Nunnally, JC. (1978). *Psychometric theory*. New York: McGraw-Hill.
Pascarella, ET, & Terenzini, PT. (1991). *How college affects students*. San Francisco: Jossey-Bass.
Pascarella, ET, & Terenzini, PT. (2005). *How college affects students (Vol. 2): a third decade of research*. San Francisco: Jossey-Bass.
Rea, LM, & Parker, RA. (2014). *Designing and conducting survey research: a comprehensive guide* (4th ed.). Hoboken: Jossey-Bass.
Royce, D. (2007). *Research methods in social work* (5th ed.). Belmont: Thompson Higher Education.
Scott, SS, McGuire, JM, & Shaw, SF. (2003). Universal design for instruction: a new paradigm for adult instruction in postsecondary education. *Remedial and Special Education, 24*, 369–379. doi:10.1177/07419325030240060801.
Smith, MK, Vinson, EL, Smith, JA, Lewin, JD, & Stetzer, MR. (2014). A campus-wide study of STEM courses: new perspectives on teaching practices and perceptions. *Cell Biology Education, 13*(4), 624–635. doi:10.1187/cbe.14-06-0108.
Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research, and Evaluation, 7*(17), 137–146.
Thompson, B, & Daniel, LG. (1996). Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197–208. doi:10.1177/0013164496056002001.
Trigwell, K, & Prosser, M. (2004). Development and use of the Approaches to Teaching Inventory. *Educational Psychology Review, 16*, 409–424. doi:10.1007/s10648-004-0007-9.
Trigwell, K., Prosser, M., & Ginns, P. (2005). Phenomenographic pedagogy and a revised Approaches to Teaching Inventory. Higher Education Research and Development, 24, 349–360. doi:10.1080/07294360500284730.
Turpen, C, & Finkelstein, ND. (2009). Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. *Physical Review Special Topics—Physics Education Research, 5*(2), 1–18. doi:10.1103/PhysRevSTPER.5.020101.
U.S. General Accounting Office [GAO]. (1996). *Content analysis: a methodology for structuring and analyzing written material*. Washington, D.C: GAO/PEMD-10.3.1.
van de Vijver, F. (2001). The evolution of cross-cultural research methods. In DR Matsumoto (Ed.), *The handbook of culture and psychology* (pp. 77–94). New York: Oxford University Press.
Walter, EM, Beach, AL, Henderson, C, & Williams, CT. (2014). *Measuring post-secondary teaching practices and departmental climate: the development of two new surveys*. Indianapolis: Paper presented at the Transforming Institutions: 21st Century Undergraduate STEM Education Conference.
Weber, RP. (1990). *Basic content analysis* (2nd ed.). Newbury Park: Sage.
Wieman, C, & Gilbert, S. (2014). The Teaching Practices Inventory: a new tool for characterizing college and university teaching in mathematics and science. *CBE-Life Sciences Education, 13*, 552–569. doi:10.1187/cbe.14-02-0023.
Zieffler, A., Park, J., Delmas, R., Bjornsdottir, A. (2012). The Statistics Teaching Inventory: a survey of statistics teachers' classrooms practices and beliefs. *Journal of Statistics Education. 20*(1). Retrieved from http://www.amstat.org/publications/jse/v20n1/zieffler.pdf