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.

Approaches to Teaching Inventory (ATI)\textsuperscript{a}

*Background.* The Approaches to Teaching Inventory (ATI, Trigwell & Prosser 2004) was created to address a lack of research into the diversity of teaching approaches at the university level. The items were generated by a qualitative phenomenographic analysis of interview transcripts from 24 university instructors. The instructors provided descriptions of their teaching approaches in response to specific scenarios. Five general categories of teaching practices were created from analysis of the interview data, including teacher-focused information transmission, teacher-focused interaction, student-teacher interaction, concept acquisition, and student focused conceptual development/conceptual change. These general categories informed the item generation process for the instrument.

*Intended Population.* All post-secondary level instructors can complete the ATI.

*Reliability and Validity.* Face validity of the ATI was established through an iterative process of interviews and pilot testing with post-secondary instructors. The instrument authors began with 74 statements representing the 5 aforementioned categories of teaching practices: information transmission, teacher focused and student-teacher interaction concept acquisition, and student focused conceptual development/conceptual change. This process revealed repetitive items, reducing the overall item number to 49. The statements were then taken to 11 of the instructors from the interviews for a pilot test. The responses from the instructors were used to cut 10 more items. This 39-item version was again piloted with 58 post-secondary physics and chemistry instructors. Based on principal component analysis and construct validity data, the instrument was reduced to the final 16 item two-construct version. Follow up research using the 16-item ATI, including confirmatory factor analysis, has likewise confirmed a two-construct structure (Trigwell & Prosser 2004; Prosser & Trigwell 2006). Furthermore, earlier work by Trigwell, Prosser, and Waterhouse (1999) triangulated the validity of a self-report instrument by comparing the self-reported teaching practices provided by an early version of the ATI to the teaching practices reported by students, providing evidence that ATI data was in agreement with student self-report.

*Respondent and Administrative Burden.* The ATI has 16 total items (9 of which we coded as teaching practice items) and no estimated time-to-completion. Items are presented in a single block in which participants indicate how true each statement is for their teaching (5-point scale: Only rarely, Sometimes, About half the time, Frequently, Always).

*Scoring Convention.* All items on the ATI are worded positively and are scored from 1 to 5, with higher scores assigned to higher frequencies. The 2004 version of the ATI has
two constructs with 8 items each, Conceptual Change/Student-focused (CCSF) and Information Transmission/Teacher-focused (ITTF). Although CCSF and ITTF scales can be considered separately, Trigwell and Prosser (2004) recommend the scales be used together for correlational studies.

Reported Analyses. ATI data has been used to correlate approaches to teaching and other factors related to the teaching and learning context such student approaches to learning, enjoyment of teaching, and perceptions of leadership (Trigwell et al. 1998; Trigwell et al. 1999; Trigwell, 2002; Gibbs & Coffey 2004).

Practices Elicited. Several items on the ATI refer to reflective practice (n=4). All of these items probe aspects the instructor’s orientation to teaching, i.e. their rationale for given instructional practices. For example, “I feel a lot of teaching time in this subject should be used to question students' ideas.” 5 items on the ATI refer to instructional format, including discussion (n=3), student reflection (n=1), and students applying or extending course concepts (n=1). The remaining items referred to teaching beliefs (n=7) (e.g. “I feel that it is better for students in this subject to generate their own notes rather than always copy mine”).

Borrego Engineering Faculty Survey (BEFS*)

Background. The BEFS (Borrego, Cutler, Prince, Henderson, & Froyd 2013) was developed in response to a growing need to document use of research-based instructional strategies (RBIS) in STEM fields. The survey elicits use of specific RBIS and core components of RBIS. The instructional strategies included on the BEFS were selected based on a review of the literature, and only strategies with documented use in engineering contexts with positive student learning outcomes were included. Similarly, the core components of RBIS were taken from the literature. A team of physics and engineering education experts (the authors) decided upon the final list of core components, many of which spanned multiple RBIS.

Intended Population. The BEFS was intended for post-secondary engineering faculty. For their study, the authors surveyed engineering faculty who had recently taught introductory level courses.

Reliability and Validity. Reliability of the BEF S was based on multiple tests of Cronbach’s alpha for internal consistency. The initial version of BEFS was created based on an earlier survey of introductory physics faculty (Henderson & Dancy 2009). Based on the results of this initial version (V1 Cronbach’s $\alpha = 0.855$), three RBIS and three critical components were removed (V2 Cronbach’s $\alpha = 0.921$). Content validity of the instrument was achieved by aligning survey concepts with those in the literature and by expert panel review (engineering and physics education faculty). The authors report no other methods to improve reliability and validity.

Respondent and Administrative Burden. The authors did not provide information on the number of items on the instrument nor an estimated time-to-completion. The BEFS uses two different response scales, one for ‘specific RBIS’ items and one for ‘critical
components’ items. ‘Specific RBIS items’ use a 5-point scale to measure use and awareness (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’ve heard the name, but do not know much else about it; I have never heard of it). ‘Critical components’ items use a 5-point proportion scale to measure the amount of time in class spent on each component (0%; 1-25%; 26-50%; 51-75%; 76-100%).

Scoring Convention. The authors scored only the participants who responded, “I currently use all or part of it” as users of an RBIS. All other responses were considered non-users. This effectively made a binary scale for RBIS use. All participants who responded above 0% on Critical Components items were considered to have spent time on the activity.

Reported Analyses. The authors reported BEFS results as response trends (percentages). Chi-square and Fisher’s exact test were used to compare use of RBIS and use of critical components.

Practices Elicited. There are two blocks of items on the BEFS. The first block asks instructors about their use of 11 different RBIS. We coded 9 of these items into instructional format codes, including: group work (n=3), group problem solving (n=2), problem solving (n=1), discussion (n=1), real world context (n=1), and students conducting a scientific investigation (n=1). The remaining 2 RBIS items were coded as assessment practice.

The second block of items examines the amount of time instructors spend in class on ‘critical components’ of each instructional strategy. Of the 16 ‘critical components’ items, we coded 10 as instructional format and 6 as assessment practice. The most common type of instructional format was providing students with a real world context (n=3). Other instructional format items found in the critical component section were each sorted into a unique code, including students applying and extending course content (n=1), real time polling of students (n=1), students reflecting on course content (n=1), students explaining course concepts (n=1), and group work (n=1). Among critical components related to assessment practice, grading policy was the most common (n=2). There were also single items related to student presentations (n=1), the nature of summative assessment questions (n=1), the nature of feedback given to students (n=1), and formative assessment (n=1).

Faculty Survey of Student Engagement (FSSE)

Background. The FSSE is a proprietary survey designed by the Center for Post-secondary Research at Indiana University (CPRIU 2012) as a companion instrument to the National Survey of Student Engagement (NSSE). Both the FSSE and NSSE were developed out of growing concern that previous measures of undergraduate education (e.g. admissions standards, faculty research prestige) were inadequate to predict student retention. Instead, researchers noted that positive student engagement with an institution was important to retention (Umbach & Wawrzynski 2005). This led to several measures being developed to address student engagement. However, faculty also play a large role in student engagement. Additional measures were therefore developed to determine the influence of faculty, one of which was the FSSE (Umbach & Wawrzynski 2005). In addition to many other elements, the FSSE provides data on post-secondary faculty members’ perspectives
on students’ relationships with campus staff, faculty-student interactions, and the perceived importance of various modes of learning.

**Intended Population.** The intended population for the FSSE instrument is all post-secondary instructors in the United States.

**Reliability and Validity.** Although the reliability and validity of NSSE is well documented in the literature and materials from the authors (e.g. Kuh, 2001), there is little published information on the reliability and validity of the FSSE. The FSSE was originally pilot tested in 2003 and has been administered annually since 2004. The survey underwent major revision in 2013, with scale refinement, improved clarity, and updated language related to technology. Factor analyses indicated 9 constructs for the 2013 FSSE (N=18,133), each with Cronbach’s alpha of 0.7 or higher (BrckaLorenz, Chiang, & Laird, 2014).

**Respondent and Administrative Burden.** The FSSE has 130 items (23 teaching practice, 12 demographic, and 95 other items) and takes 15-20 minutes to complete. The teaching practice items on the FSSE are presented in three blocks, one 8-item block of instructional format items and two 12-item blocks, each a mix of instructional format and reflective practice items. The 8-item block uses an 8-point scale referring to the amount of time spent in class on a given practice (0%; 1-9%; 10-19%; 20-29%; 30-39%; 40-49%; 50-74%; 75% or more). The two 12-item blocks use a 4-point response scale (Very important, Important, Somewhat Important, Not Important).

**Reported Analyses.** FSSE data are reported in frequency distributions that provide counts and percentages for how faculty responded. For each item, the number and percentage of faculty is provided for each possible answer choice (FSSE, 2012).

**Practices Elicited.** The majority of the FSSE teaching practice items (n=13) were coded as instructional format. The most common instructional format code was students applying and extending course content (n=3). Other instructional format types included instructor demonstration (n=1), discussion (n=1), lecture (n=1), group work (n=1), lab or experiment activities (n=1), and students explaining ideas or course content to each other (n=1). In addition, 4 items were coded as vague instructional practice. Seven items were coded as reflective practice. The majority of these elicited how often faculty encouraged students to engage in strategies for success in their course (n=6), including students asking other students for help with course content and students reviewing notes after class. The remaining reflective practice item referred to instructional goals. Lastly, 3 items were coded as assessment practice. These items covered the nature of feedback given to students (n=1), student presentations (n=1), and group assessments (n=1).

**Henderson and Dancy Physics Faculty Survey (HDPFS*)**

**Background.** The HDPFS (Henderson & Dancy 2009; Henderson, Dancy, & Niewiadomska-Bugaj 2012) was designed to assess knowledge and use of research based instructional strategies (RBIS) in post-secondary physics instruction. There have been
many such strategies developed in response to calls for reform in introductory physics teaching. The authors selected the instructional strategies for the HDPFS based on those identified in the physics education literature with documented positive effects on student learning outcomes.

**Intended Population.** The intended population for the HDPFS is college physics faculty teaching introductory quantitative physics courses (algebra- or calculus-based). These courses were selected as the intended context because they have the highest enrollment and the majority of physics RBIS have been developed with these courses in mind.

**Reliability and Validity.** The authors addressed content validity for the HDPFS by creating the HDPFS in collaboration with experts at the American Institute of Physics Statistical Research Center. No other validity or reliability measures were reported.

**Respondent and Administrative Burden.** There are 61 items on the HDPFS (40 teaching practice, 10 demographic, and 11 other items) and no estimated time-to-completion. The items describing RBIS ask participants to “provide their level of familiarity” with each RBIS using a 5-point scale (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’ve heard the name, but do not know much else about it; I have never heard of it). The HSPFS also has follow up questions to the RBIS items. These questions elicit how much the participants “had modified” or “how long they had been using” the RBIS.

HDPFS response scales differ for instructional format items versus assessment items. Instructional format items ask participants to respond on a 6-point frequency scale (Never; Once or twice; Several Times; Weekly; For Nearly Every Class; Multiple Times Every Class). Assessment items ask participants to describe their use of assessment question formats using a 4-point frequency scale (Never Used on Tests; Used Occasionally on Tests; Used Frequently on Tests; Used on All Tests).

**Reported Analyses.** The authors reported HDPFS response trends as percentages.

**Practices Elicited.** There are three blocks of items on the HDPFS: specific named RBIS, general teaching practices, and assessment practices. The specific named RBIS section was intended to determine participants’ knowledge and use of 24 RBIS. The named RBIS items include items on problem solving \((n=5)\), lab or experiment \((n=3)\), group work \((n=1)\), group problem solving \((n=1)\), formative assessment \((n=1)\), and demonstration \((n=1)\). The 14 remaining RBIS items were coded as vague instructional practice. The general teaching practice items were coded under instructional format and included group work \((n=2)\), problem solving \((n=2)\), lecture \((n=1)\), demonstration \((n=1)\), and real time polling of students \((n=1)\). The majority of the assessment practices items \((n=6)\) were related to the nature of the assessment questions. The other assessment practice item was coded as vague assessment practice.
Higher Education Research Institute (HERI) Faculty Survey

Background. The Higher Education Research Institute (HERI) faculty survey (Hurtado, Eagan, Pryor, Whang, & Tran 2012) measures multiple aspects of faculty life, such as faculty student interactions, preferred teaching methods, and perceptions of institutional climate. A national-level and proprietary survey, it was originally created to evaluate national general education programs funded by the Exxon Education Foundation. HERI has been administered every three years since 1990, with the most recent and available version administered in 2011.

Intended Population. The HERI faculty survey is intended for post-secondary instructors in the United States.

Reliability and Validity. Hurtado et al. (2012) tested the 11 HERI constructs through a multi-step construct validity process. They first created an initial item pool based on existing HERI faculty survey items. Selection of the factors was based in the literature. The authors then conducted an exploratory factor analysis on the constructs, resulting in a final categorization of HERI items into one of 11 constructs (Sharkness, DeAngelo, & Pryor, 2010).

Respondent and Administrative Burden. There are 284 items on the HERI, (35 teaching practice, 20 demographic, and 229 other items). The instrument takes approximately 25 minutes to complete. A portion of the instructional format items and assessment practice items ask participants the “number of courses they use each practice” on a 4-point scale (All, Most, Some, None). The reflective practice items and the remaining instructional format items ask participants “how often they encourage a practice” on a 3-point scale (Frequently, Occasionally, Not at all).

Scoring Convention. Construct scores are calculated using Item Response Theory. Scores are estimated based on the pattern of responses given by the participant for the entire set of items on a given construct (Sharkness et al. 2010). Scores are calculated on a Z-score metric and rescaled to a mean of 50 with standard deviations of 10. For the purposes of reporting, participants are sorted into low, average, and high scoring groups. Low scoring participants are those with scores one half standard deviation below the mean score. Average are those participants within one half standard deviation of the mean score. The high score group are those participants that are at least one half standard deviation above the mean score.

Reported Analyses. HERI results are reported in frequency distributions. Percentages of the participants are reported for each response choice for each item. For the constructs, percentages of participants are reported for each score group (low, average, high) for each construct (Hurtado et al. 2012).

Practices Elicited. There are 35 teaching practice items on the 2011 HERI, most of which refer to instructional format (n=15) or assessment practice (n=13). We coded the remaining items as reflective practice (n=7). The most common instructional format items are: students explaining course ideas or concepts (n=2), students having input in
course content \((n=2)\), and providing a real world context \((n=2)\). The most common assessment practices items include: nature of assessment questions \((n=3)\), writing assignments \((n=3)\), and grading policy \((n=2)\). Most of the reflective practice items focused on how often instructors encouraged strategies for student success \((n=6)\), e.g. encouraging students to revise papers, seek alternative solutions to problems, and use scientific research articles (Hurtado et al. 2012).

**National Study of Post-secondary Faculty (NSOPF)**

**Background.** The National Study of Post-secondary Faculty (NSOPF) survey was intended to be the primary source of data for a long-term cross-sectional study (NCES 2004; Heuer et al. 2006). The NSOPF project was conducted by RTI International and funded by the National Center for Education Statistics at the U.S. Department of Education. The most recent version (2004) was based on three previous iterations, designed to maintain continuity with previous versions while addressing emerging issues relevant to post-secondary instructors (e.g. online education). Items on the NSOPF elicit a wide variety of faculty characteristics including professional background, scholarly activities, and incentives.

**Intended Population.** The intended population for the NSOPF is all post-secondary faculty in the United States.

**Reliability and Validity.** The NSOPF authors had two primary goals during their development process, maintaining items from previous versions of the instrument for trend analysis and adding items that probe emerging issues. The most recent version (2004) of the NSOPF was primarily based off of the 1999 version. Suggestions from the project Technical Review Panel, respondents for the 1999 survey, government officials, and post-secondary researchers were incorporated to improve the survey. These groups of individuals then reviewed a field test version of the instrument. Pilot data was subsequently collected using the revised field test instrument. The pilot data allowed the research team to evaluate the field test instrument in an environment similar to that of the final instrument.

**Respondent and Administrative Burden.** There are 83 items on the NSOPF (10 teaching practice, 36 demographic items, 37 other items). The average time-to-completion is 30 minutes. The 10 teaching practices items on the NSOPF are presented in the same block. Participants respond to the items using a 3-point scale (Used in all classes, Used in some classes, Not used; NSOPF 2004).

**Reported Analyses.** Results of the NSOPF are reported by the number of responses and by unweighted and weighted percentages. The NSOPF authors recommend several statistical packages for analyzing NSOPF data including SUDAAN, WesVar, Stata, and SAS. These packages are capable of accounting for its stratified two-stage sampling design (Heur et al. 2006, pp. 97-99).

**Practices Elicited.** Most of the instructional practice items on the NSOPF describe assessment practice \((n=9)\). Assessment practices items include the nature of questions on
summative assessments \((n=3)\), writing assignments \((n=2)\), student presentations, group assessments, and peer evaluation of student work. There is one instructional format item related to using a real world context.

**On the Cutting Edge Survey (OCES)**

*Background.* The On the Cutting Edge (OCES) survey (MacDonald, Manduca, Mogk, & Tewksbury 2005) was developed as part of an NSF-funded program for geoscience faculty professional development. The authors recognized a lack of available research on instructional practices in the geosciences, despite increased focus on instructional practices in other disciplines. The OCES was intended to provide a national snapshot of teaching practices in undergraduate geosciences that could be used as a baseline for future professional development programs and to encourage faculty to improve their practice.

*Intended Population.* The OCES is intended to sample a nationwide group of undergraduate geosciences faculty, including individuals at public and private 2- and 4-year institutions.

*Reliability and Validity.* Extensive literature review, pilot testing, interviews, and corroboration of the authors and experts functioned to improve the validity of the OCES. The instrument was developed with input from the Statistical Research Center of the American Institute of Physics. The authors relied on a previous instrument from the biological sciences and *Using Data in Undergraduate Classrooms* (Manduca & Mogk 2003) to identify teaching practices to elicit in the survey. Face validity was addressed by interviews with five faculty and a pilot test with 16 faculty at an American Geophysical Union meeting.

*Respondent and Administrative Burden.* The OCES has 46 items (29 teaching practice, 7 demographic, and 10 other items) and no cited time-to-completion. The survey has 4 main blocks, 1 assessment practice block and 3 instructional format blocks. The assessment practice items ask participants to respond using a binary scale (check boxes). The first block of instructional format items asks participants to rank their level of use of in-class activities on a 5-point frequency scale (Never; Once or twice; Several times; Weekly; For nearly every class). The second is a series of student problem solving items that have participants respond on a binary scale (check boxes). For the final instructional practice block asks participants rate a set of student problem solving items on a 5-point frequency scale (Never; Once or twice; Several times; Weekly; For nearly every class).

*Reported Analyses.* Results were reported as frequency counts and percentages.

*Practices Elicited.* Most items on the OCES refer to instructional format \((n=22)\). The most common instructional practices elicited by the OCES refer to students analyzing data sets \((n=5)\), problem solving \((n=5)\), the instructor providing a real world context \((n=2)\), instructor-led question and answer \((n=2)\), and group work \((n=2)\). There are 7 additional items related to assessment, including: writing assignments \((n=1)\), student presentations \((n=1)\), grading policy \((n=1)\), and peer evaluations of student work \((n=1)\).
The three remaining assessment items were coded as vague assessment practices.

**Post-secondary Instructional Practices Survey (PIPS)**

*Background.* The PIPS (Walter et al. 2014) was developed out of the growing need to measure the impact of change initiatives for improving STEM teaching practices. After reviewing existing instruments, the authors determined that a concise, interdisciplinary, and non-proprietary instrument was needed for measuring faculty self-report of teaching practices. A self-report instrument was chosen because the PIPS authors argue that self-report instruments are easier to implement, cheaper than other methods such as observational protocols, and capable of capturing practices that are difficult to observe.

*Intended Population.* The intended population for the PIPS is all post-secondary instructors.

*Reliability and Validity.* The initial PIPS items were created based on concepts and items from 4 existing instructional practice instruments (ATI, FSSE, HERI, NSOPF) and 2 observational protocols (Teaching Dimensions Observational Protocol, Hora, Oleson, & Ferrare 2012; Reformed Teaching Observational Protocol, Piburn et al. 2000). The PIPS authors also drew upon 4 literature reviews (Iverson 2011; Meltzer & Thornton 2012; Pascarella & Terenzini 1991;2005). The PIPS authors reduced the initial item pool \(n=153\) through multiple rounds of item removal, revision and generation. The items were revised for clarity by 4 external educational researchers. The result of the item revision process was 24 teaching practice items. These items were pilot tested with 827 post-secondary instructors from 4 higher education institutions. Exploratory and confirmatory factor analyses were completed resulting in a 2-factor and 5-factor solution being supported. The 2-factor solution includes factors for student-centered practice and instructor-centered practice. The 5-factor solution includes factors for student-student interactions, content delivery, formative assessment, student-content engagement, and summative assessment. Additionally, the overall instrument reliability was calculated using Cronbach’s Alpha \(\alpha=.800\).

*Respondent and Administrative Burden.* There are 33 total items (24 teaching practice items, 9 demographic items) on the PIPS. The reported estimated time-to-completion is 10 minutes. The teaching practice items are presented in a single block. All teaching practice items use the same 5-point scale (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).

*Scoring Convention.* The scoring convention developed by the PIPS authors scores each item from 0-4. Higher scores are assigned to responses indicating greater descriptiveness of ones teaching. Scores are calculated by construct by adding the scores for all items in a construct, dividing by the total possible points for the construct, and multiplying by one hundred. Scores are reported from 0-100.

*Reported Analyses.* ANOVA and independent T-tests were used to compare demographic
groups based on their average PIPS scores. Correlational analysis was also used to examine relationships between PIPS factors and aspects of teaching such as class size, years teaching, and time spent in lecture.

**Practices Elicited.** The items included on the PIPS consist of instructional practice items \( (n=11) \), assessment practice items \( (n=9) \), and reflective practice items \( (n=4) \). The instructional practice items include items related to discussion \( (n=3) \), students explaining ideas and concepts \( (n=1) \), conceptual frameworks \( (n=1) \), lecture \( (n=1) \), and instructor led question and answer \( (n=1) \). Additionally, there are instructional practice items concerning the instructor providing a real world context \( (n=1) \), student reflection \( (n=1) \), students having input \( (n=1) \), and group work \( (n=1) \). The assessment practice items were primarily related to formative assessment \( (n=2) \), grading policy \( (n=2) \), and the nature of assessment questions \( (n=2) \). There were also assessment items coded under assessment of specific outcomes \( (n=1) \), the nature of feedback to students \( (n=1) \), and generic assessment \( (n=1) \). The reflective practice items consist of the instructors’ orientation toward teaching \( (n=3) \) and instructional goals \( (n=1) \).

**Survey of Teaching Beliefs and Practices (STEP)**

**Background.** The STEP (Marbach-Ad, Schaefer-Zimmer, Orgler, Benson, & Thompson 2012) was developed to determine the impact of a teaching and learning center on the teaching practices of the faculty at the authors’ institution. The center was created in response to frequent calls in the literature for improvements in undergraduate science education. It was primarily designed to promote student engagement techniques of chemistry and biology faculty. After the center had been active for 5 years, the authors developed surveys for faculty (the STEP), graduate students, and undergraduate students to determine how the center was affecting the teaching practices of faculty.

**Intended Population.** The intended population for the STEP is chemistry and biology faculty who participated in the teaching and learning center at the authors’ institution.

**Reliability and Validity.** An iterative review process was used to develop the STEP. Review by a diverse group of experts and pilot testing were used to establish face and content validity. Science, education, and psychology experts, including department chairs, faculty members, and external evaluators, reviewed the STEP for content and face validity. Graduate students and a statistician also reviewed the survey. No other validity or reliability measures were reported.

**Respondent and Administrative Burden.** The STEP survey has 86 items (30 teaching practice, 8 demographic, and 48 other items) and no estimated time-to-completion. Items are organized into ‘use of teaching practice’ and ‘perceptions of teaching practice’ blocks. ‘Use of teaching practice’ items on the STEP use different response scales for instructional format and assessment practice items. Instructional format items use a 5-point scale (Not Used; Once Per Semester; A Few Times a Semester; Most Class Sessions; Almost Every Class Session) and the assessment practice items use a 3-point (Do Not Use; Use and Counts Towards Student Grade; Use and Does Not Count Towards
Student Grade). The ‘perceptions of teaching’ items on the STEP use a 5-point scale (Not Important; Slightly Important; Fairly Important; Important; Very Important).

Scoring Convention. All of the 5-point scale items on the STEP (instructional format and perceptions of teaching) are scored from 1 to 5, with higher scores assigned to responses corresponding with either greater importance or more frequent use. No scoring convention was reported for the assessment items.

Reported Analyses. The authors used t-tests to make comparisons between mean scores on an item-by-item basis and reported response trends as percentages.

Practices Elicited. The items include both instructional format (n=16) and assessment items (n=12). The most common instructional format items described group work (n=3) and discussion (n=1). Other instructional format item types include instructor led question and answer (n=1), problem solving (n=1), students reading the primary literature (n=1), conceptual framework (n=1), making connections to scientific research (n=1), students creating and or analyzing data sets (n=1), and providing a real world context (n=1). The remaining 3 instructional format items were coded as vague instructional formats. For assessment practices, items related to the nature of assessment questions (n=2) and writing assignments (n=2) were most common. Other assessment practice item types are student presentations (n=1), peer evaluation of student work (n=1), and formative assessment (n=1).

Statistics Teaching Inventory (STI)

Background. The Statistics Teaching Inventory (STI) (Zieffler, Park, Delmas, & Bjornsdottir 2012) was created in response to new instructional and assessment guidelines released by the American Statistical Association in 2005. These guidelines focused on promoting statistics literacy through the use of real data, stressing conceptual understanding, incorporating technology, and aligning assessments with course goals. The guidelines were well received by the statistics education community and led to the development of workshops, publications, and textbooks aligned with the guidelines. The authors developed items on the STI using the American Statistical Association guidelines in order to assess the impact of the guidelines on the teaching of statistics instructors.

Intended Population. The intended population for the STI is post-secondary and advanced placement secondary statistics instructors.

Reliability and Validity. The first version of the instrument of the STI had 102 items. A panel of statistics education community members and NSF project advisors to establish content validity. Face validity of the instrument was established through think aloud interviews, including two groups of post-secondary faculty teaching outside of statistics and mathematics and a group of post-secondary statistics faculty from across the United States. During these interviews, participants were read the items and asked to articulate their responses and thought processes. The results of the interviews and a second small pilot with local statistics educators were used to create the final 50-item version of the instrument.
**Respondent and Administrative Burden.** The STI has 50 total items (21 teaching practice, 7 demographic, and 22 other items) and no estimated time-to-completion. It has 3 different response scales. Instructional format items on the STI use a 5-point frequency scale (Never; Seldom; Some of the Time; Most of the Time; All of the Time). Assessment practice items use a binary scale (Agree/Disagree). Teaching and assessment beliefs items use a 4-point agreement scale (Strongly Disagree; Disagree; Agree; Strongly Agree) with an “undecided” option.

**Scoring Convention.** A scoring scheme was developed in order to generate an overall STI score. The response choices for each item were assigned a score between 0 and 1. Higher scores were given to responses corresponding with more reform-oriented practices and beliefs. No other specialized data analysis techniques were reported.

**Reported Analyses.** Results were reported as response trends (percentages) and average response levels for specific items.

**Practices Elicited.** The STI contains primarily instructional format and assessment practice items. Although the STI includes statements that elicit teaching and assessment beliefs, these were excluded from analysis, as they did not imply a direct connection to teaching practice, e.g. “Rules of probability should be included in an introductory statistics course.” The instructional format items (n=10) include: students analyzing data sets (n=3), lecture (n=1), discussion (n=1), problem solving (n=1), and instructor demonstration (n=1). The three remaining instructional format items were coded as vague instructional format items. The assessment practice items (n=9) were primarily written to elicit how summative assessments assess specific outcomes (n=6), for example: “My assessments evaluate students’ ability to critically examine statistics in the media.” The assessment practice items also included one ‘nature of summative assessment questions’ item and one item on group assessments.

**SUCCEED Survey**

**Background.** The Southeastern University and College Coalition for Engineering EDucation (SUCCEED) survey (Brawner, Felder, Allen, & Brent 2001; 2002) was created to evaluate teaching professional development programs for engineers developed by the SUCCEED coalition, a group of NSF-funded Engineering Education Coalition schools in the southeastern United States. The goal of SUCCEED was to develop and disseminate new instructional methods, materials, and programs to improve learning of engineering students. The survey covers five main areas, including non-classroom teaching experience, perceived importance of teaching, frequency of use of teaching practices, involvement in teaching improvement programs, and use of technology for instruction (including the internet and email). The instrument authors used SUCCEED data to determine the effectiveness of related professional development for changing the teaching practices of faculty.

**Intended Population.** The intended population for SUCCEED is post-secondary engineering faculty at one of 8 Engineering Education coalition schools.
Reliability and Validity. The SUCCEED survey was administered in 1997, 2002, and 2003. Item wording was modified slightly between implementations to improve clarity. In addition, a pilot test was conducted before the survey was first administered in 1997, which led to minor item revisions. There were no other methods reported to establish reliability and validity.

Respondent and Administrative Burden. SUCCEED has 67 total items (12 teaching practice items, 7 demographic items, 48 other items) and no estimated time-to-completion. Nine of the teaching practice items are in one block. These items are a mix of instructional format and assessment practices items. Participants are asked to respond on a 5-point scale (Every class, One or more times a week, One or more times a month, One or more times a semester, Never). There is also one assessment practice item related to group assessment. This item has participants respond on a 3-point scaled (In every course I teach; In some but not all courses I teach; Never). In addition, 2 items related to instructional objectives have a 4-point response scale (Always; Usually; Sometimes; Never).

Scoring Convention. The survey authors scored the teaching practice items on SUCCEED from 0 to 4, with the exception of the solitary assessment practice items scored 0 to 2. Higher scores were assigned to higher frequency. Similarly, the instructional objectives items were scored from 0 to 3, with higher scores assigned to higher frequency.

Reported Analyses. One-way ANOVA with Bonferroni tests were used to compare mean scores.

Practices Elicited. The SUCCEED survey includes instructional format (n=5), assessment practice (n=5), and reflective practice items (n=2). The instructional format items cover group work (n=2), lecture, instructor demonstration, and instructor-led question and answer. Assessment practices include group assessments (n=3) and writing assignments. There are also 2 reflective practice items aimed at instructional goals.

Teaching Practices Inventory (TPI)

Background. The Teaching Practices Inventory (TPI, Wieman & Gilbert 2014) was developed in response to calls for the adoption of research based teaching practices in post-secondary STEM disciplines. The TPI authors argue that a major difficulty in changing post-secondary STEM instructors teaching practices is that many universities do not measure the teaching practices that are being used in their courses. The authors claim is supported by polling of the Association of American Universities (AAU), the American Public and Land Grant Universities, and the Association of American Colleges and Universities which found that no institutions reported collecting data on teaching practices currently being used in their courses. Outside of observation protocols, the TPI authors were unaware of any method for efficiently and consistently measuring teaching practices. The TPI authors argue that observation protocols do not allow for efficient collection of large data sets because they require extensive training and observations.
Therefore, the TPI was developed to measure a broad range of teaching practices efficiently.

**Intended Population.** The intended population for the TPI is post-secondary science and mathematics instructors.

**Reliability and Validity.** The authors of the TPI did not use traditional statistical measurements of reliability and validity. They argue that their instrument serves as a rubric or checklist of possible research supported and non-research supported teaching practices for a given course. As a result, the authors state there is no underlying construct being measured, making statistical analysis of the item loadings inappropriate. The initial items on the TPI were developed based on the experience of the authors, their knowledge of the science education literature, and 7 teaching practice factors in the PULSE Vision and Change Course Level Rubric (PULSE, 2013). Two pilot tests of 150 and 179 post-secondary instructors respectively were conducted with the TPI. These pilot tests resulted in the addition of some items and response choices to represent teaching practices missed by the initial items. In addition, the instrument went through multiple rounds of expert review in between the two pilot tests.

**Respondent and Administrative Burden.** The TPI consists of 72 total items (60 teaching practice items, 12 other items). The average time to complete the instrument was 13 minutes. The teaching practice items are organized in 7 sections. The majority of the teaching practice items have a binary response scale in the form of check boxes. There is one item related to the portion of class time spent lecturing that uses a 5-point percentage scale (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Another item related to the time spent on the process by which a theory or model is developed uses a 3-point percentage scale (0-10%, 10-25%, more than 25%). There are also several items that ask respondents to fill in the number of instances per class or a percentage.

**Scoring Convention.** A scoring scheme is included in the manuscript for the TPI (Wieman & Gilbert 2014). Scores between 1 and 3 were assigned to each of the teaching practice items based on their potential for improving student learning in science and mathematics. Scores were assigned based on the judgment of the authors, support from the literature, and expert review. The majority of the items were assigned points based on direct evidence for their effectiveness from the science and mathematics education literature. Four of the items, which are related to instructors using inventories or instruments to measure student learning, are not directly supported in the science and mathematics education literature. However, the authors point there is indirect support for using these instruments in the cognitive psychology literature. Higher scores on the TPI indicate instructors who use more research based teaching practices.

**Reported Analyses.** Mean scores and standard deviation were reported by department along with average scores weighted by course enrollment. Average scores were also reported for each section of the instrument.
Practices Elicited. The TPI includes instructional format ($n=25$), assessment practice ($n=29$), and reflective practice items ($n=6$). The instructional format items consist of real time polling ($n=7$), demonstration/example ($n=4$), discussion ($n=3$), reflection ($n=2$), students have input ($n=2$), and generic instructional practice ($n=2$). In addition, there are items related to connecting course content to scientific research ($n=1$), lecture ($n=1$), question and answer ($n=1$), students reading the primary literature ($n=1$), and the instructor providing a real world context ($n=1$). The assessment practice items were coded under grading policy ($n=13$), formative assessment ($n=7$), and generic assessment ($n=5$). There were also assessment practice items associated with group projects ($n=1$), nature of feedback to students ($n=1$), student presentations ($n=1$), and writing assignments ($n=1$). The reflective practice items were related to instructional goals ($n=4$) and instructors encouraging their students to engage in strategies for success ($n=2$).

References
Heuer, R., Kuhr, B., Fahimi, M., Curtin, T. R., Hinsdale, M., Carley-Baxter, L., & Green,


