2014

Development of a Student Self-Reported Instrument to Assess Course Reform

R.C. Morris
Purdue University

Loran Carleton Parker
Purdue University

David Nelson
Purdue University

Matthew D. Pistilli
Purdue University

Adam Hagen
Purdue University

See next page for additional authors

Authors
R.C. Morris, Loran Carleton Parker, David Nelson, Matthew D. Pistilli, Adam Hagen, Chantal Levesque-Bristol, and Gabriela Weaver

Educational Assessment, 19:302–320, 2014
Published with license by Taylor & Francis Group
ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2014.964119

Routledge
Taylor & Francis Group

Development of a Student Self-Reported Instrument to Assess Course Reform

R. C. Morris, Loran Carleton Parker, David Nelson, Matthew D. Pistilli, Adam Hagen, Chantal Levesque-Bristol, and Gabriela Weaver
Purdue University

This study examines the development and implementation of a survey-based instrument assessing the effectiveness of a course redesign initiative focused on student centeredness at a large midwestern university in the United States. Given the scope of the reform initiative under investigation in this study, researchers developed an instrument called the Classroom Experience Questionnaire (CEQ), which was administered to students enrolled in redesigned courses. Early findings demonstrate strong construct validity and internal reliability of the CEQ instrument as well as concurrent validity between the CEQ and observation data gathered in concert with self-report data. The authors conclude that in the absence of trained classroom observers, the developed student self-report protocol can serve as a useful tool for measuring the constructivist orientation of pedagogy and student-centered nature of the learning environment in a higher education setting.

INTRODUCTION

Calls for teaching and curriculum reforms in higher education, particularly as they relate to student learning outcomes, are not new (McCray, DeHaan, & Schuck, 2003; Miller & Groccia, 2011, p. 102). Reform initiatives, many of which get initiated or guided by a regional accrediting agency, seek to improve the quality and level of student learning (Provezis, 2010). At present, large, research-intensive universities are the target of calls for teaching and curriculum reform. In the book *Academically Adrift*, Arum and Roksa (2010) implied that higher education is generally in need of curriculum change to ensure that postsecondary matriculation provides students with the higher order cognitive skills they need to be successful. In particular, declining graduation rates in science, technology, engineering, and mathematics (STEM) disciplines, particularly for women and minorities, have spurred efforts in the United States to fix what

Correspondence should be sent to R. C. Morris, Department of Sociology, Purdue University, 700 West State Street, West Lafayette, LA 47907. E-mail: rmorris@purdue.edu

has been described as a "leaky pipeline" (Astin & Astin, 1992; Blickenstaff, 2005; De Welde, Laursen, & Thiry, 2007; Dean & Fleckenstein, 2007; Potočnik, 2009; Watson & Froyd, 2007; Zimpher, 2009). The National Science Foundation (2012) recently reported that "undergraduate attrition out of agricultural/biological sciences, mathematics/physical/computer sciences, and engineering is greater than transfers into those fields" (pp. 2–23).

One often proposed solution to curtail higher education attrition rates while improving higher education pedagogy involves the transformation of the classroom from the traditional instructor-centered model to a learner-centered one (Blackie, Case, & Jawitz, 2010; Carnell, 2007; Hughes, 2007). Common critiques of the traditional instructor-centered model argue that students get forced into a complacent role, passively consuming and regurgitating information, with little or no active-student engagement or deep learning of content (Freire, 2000; King, 1990, 1993). The learner-centered model encourages the instructor and learners to actively "construct" knowledge, which in turn builds "extensive cognitive structures that connect the new ideas and links them to what is already known" (King, 1993, p. 30). In this article we use the term *reform* to define a course that has been transformed through a pedagogical redesign strategy emphasizing active student engagement in a student-centered learning environment. Transforming the higher education classroom into an engaging student-centered environment is one way to enhance the higher order cognitive skills that are required for success after college (Arum & Roksa, 2010). Adding support to calls for change, growing evidence finds that reformed pedagogy does, in fact, increase student learning (cf. Cornelius-White, 2007; King, 1989, 1990; Slavin, 1995; Twigg, 2006).

The growing movement for reform within higher education as well as empirical findings supporting learning gains prompted the U.S. government to allocate a sizable portion of the 2014 U.S. fiscal year budget to support higher education reform of the learning environment as a way to address the leaky pipeline in STEM programs:

> . . . realignment of ongoing STEM education activities to **improve** the delivery, impact, and visibility of STEM efforts. . . . The Department of Education's role will include developing STEM Innovation Networks to **reform** STEM instruction and supporting a corps of Master Teachers who can serve as a national resource for ***improving*** *STEM teaching and learning*. NSF will focus on efforts to **improve** STEM undergraduate education and to **reform** graduate fellowships so they reach more students and align with national needs [emphasis added]. (Office of Management and Budget, 2013, p. 84)

> As part of this investment, i3 will also support up to $65 million for the Advanced Research Projects Agency for Education, which will *aggressively pursue technological breakthroughs that* ***transform*** *educational technology and empower teaching and learning* [emphasis added]. (Office of Management and Budget, 2013, pp. 80–81)

Calls for reform targeted at STEM programs are a popular talking point offered to stem the tide of graduation losses in these fields, but the changes needed to reform higher education curriculum and teaching practices are not unique to these programs. In fact, the attention and pedagogical development that has been aimed at STEM can be used to shape a holistic application of reform within higher education more generally (Denton, 1998).

As calls for reform in higher education continue, and as reform initiatives get developed and implemented, the collection of evidence regarding best practices for implementing these

changes becomes paramount. This is particularly true when extending some of the pioneering work done in STEM specific redesigns to a broader campus community. Typically, instructors get exposed to reformed or redesigned instructional practices and are encouraged to align their own course and teaching approaches with these models (Turner, 2009; Turner & Carriveau, 2010), but once an instructor adopts a pedagogical strategy designed for one program, how can they know if the reformed course is more engaging and student centered when used within their own program? It is difficult, if not impossible, to create a standardized measurement tool that is applicable to all reformed teaching and learning practices (Arum & Roksa, 2010). Too, the resource-intensive costs of evaluating teaching through direct observation methods steer many programs toward other data-gathering methods that are less directly connected to the learning environment (Hill et al., 2012), often leading programs to measure the efficacy of reform through faculty self-reports of (reformed) teaching practices (Kuh, 2003).

The time and expense associated with observation research requiring specialized training and staff resources is another reason that faculty self-reporting is relied upon and the reason why teaching practices are rarely examined beyond this level of analysis (Swing & Coogan, 2010). As Jenny (1996) and Swing and Coogan (2010) indicated, personnel time associated with assessment is generally not counted separately from faculty and staff members' general duties. V. B. Harper (2009) indicated that as much as 30% of one administrator's time could be allocated to the assessment of just one stated learning outcome, and as much as 50% of a faculty member's time, depending on the activities required for the assessment activity. Thus, instructional practices resulting from reform initiatives rarely get observed; and, unfortunately, when observations are completed, they tend to differ significantly from instructor self-reports (Bell et al., 2012; Ebert-May et al., 2011). Instructor self-reports have the strength of being easy to collect but a demonstrated weakness when compared to more robust observation methods.

The research reported in this study compares the assessment practices of trained and calibrated observers to a third alternative: student self-report methods (for more on using student self-report data for assessment purposes, see Centra, 1993; Ebert-May et al., 2011; Pace, 1985; Pike, 2011). The notion of using student self-report data as a potential replacement for observation methods was questioned by many involved with our redesign program. Questions raised during the initial research design stages related to the validity of student self-reported data as a measure of course reform. This concern gives root to the central research question of this article: Can student self-report data be used successfully as a supplement to, or replacement of, observation methods when assessing student centeredness and student engagement? Research conducted by Astin (1971, 1977), Kuh (2001), and Pace (1984, 1985) on large samples of undergraduate students have resulted in highly useful and reliable data. In addition, Pike (2011) noted that "60 percent of the articles published in [higher education] journals in 2006 made use of self-report data from surveys" (p. 41). Pike also indicates that a second informal study on the same journals for 2010 yielded roughly 50% of data stemming from self-perceptions of students, alumni, faculty, and staff. We use student self-report data based on this tradition, subjecting our data to a rigorous series of tests demonstrating the robustness of self-assessment as a valid and reliable methodology for rating specific elements of the learning environment (Pintrich, Smith, Garcia, & Mckeachie, 1993).

This article outlines the research context of course reform under investigation and the methods employed to assess reform, and it presents results supporting the development of a student self-reported survey instrument.

THE RESEARCH CONTEXT

This research began as an evaluation of the efficacy of a university-wide course reform project on the main campus of Purdue University. Purdue University is a large land grant university in the midwestern United States offering 200 undergraduate majors and graduate degrees in just over 70 academic programs. Purdue has a strong and long-standing reputation for STEM and for developing pedagogical strategies for STEM specific program reforms (cf. Borrego, Streveler, Miller, & Smith, 2008; Haghighi, 2005; Katehi et al., 2004; Streveler, Litzinger, Miller, & Steif, 2008). As a result of this shared focus, Purdue began a campuswide reform initiative. The Instruction Matters: Purdue Academic Course Transformation (IMPACT) project focuses on creating learner-centered courses and classrooms that make the learning process more active and engaging by improving student-centered pedagogy at this research intensive campus. Supplementing the STEM redesign background at Purdue the IMPACT project is largely inspired by the work done at the National Center for Academic Transformation (NCAT), an organization supported by foundational grants with the purpose to "produce better learning outcomes for students" ("The National Center," 2013). The NCAT focuses on cost-saving and strong learning gains (primarily through the use of technology); assessment of the NCAT reformed courses have resulted in positive changes in learning environments and statistically significant improvement in student learning for dozens of courses since the Center's inception in 1999 (Twigg, 2006).[1]

IMPACT targets large, foundational courses from all disciplines, supplementing several preexisting discipline-specific course redesign efforts. The broad scope of the IMPACT initiative requires a program fostering best practices in course redesign as well as flexibility for faculty and departments to enact reforms that meet program specific needs and contexts. IMPACT initially targeted 10 courses and has since expanded to reach 30 courses each year.

The redesign process begins for a faculty member when they join IMPACT professional development. Professional development places instructors in a supportive learning community with other instructors and education specialists. The development process focuses on supporting established learning objectives as well as the creation of new learning objectives. Emphasis gets placed on assessing student achievement of learning outcomes. At all stages, course redesign emphasizes appropriate use of innovative technological solutions and learning spaces designed to increase active learning.

HIGHER EDUCATION REFORM AND ASSESSMENT

The broad scope and size of the IMPACT reform effort necessitated a focused and efficient assessment strategy. The majority of transformation initiatives in higher education have originated from within specific departments or colleges, with most assessment measures designed to gauge course transformation from the limited pedagogical and disciplinary culture of certain fields. The predominance of STEM disciplines among course transformation initiatives has

[1]Of the 10 Research-1 or Research-2 institutions that completed the NCAT 3-year program in 2006, seven had redesigns limited to STEM disciplines (Twigg, 2006).

given rise to many assessment tools that emphasize particular applied practice of the hard sciences like physics, biology, math, and chemistry. These protocols are often geared toward observational metrics that posit a particular method for pedagogy and student experience within a classroom. Such is the case with the commonly cited Reformed Teaching Observation Protocol (RTOP; Piburn et al., 2000; Sawada et al., 2000). The RTOP requires observers to evaluate the pedagogical practices of an instructor against a set of best practice criteria that are specific to particular STEM disciplines (Sawada, Piburn, & Judson, 2002). More recent efforts like the Classroom Observation Protocol for Undergraduate STEM (Smith, Jones, Gilbert, & Wieman, 2013), the Protocol for Language Arts Teaching Observation (Grossman et al., 2010) also rely upon frequent documentation of specific instructional practices. Although such observation protocols provide detailed formative information for instructors, significant time and money is required to train observers and coordinate observations using these protocols (Meyer, Cash, & Mashburn, 2011).

The cost and limited scope of these instruments significantly limits their use as assessment tools for the large-scale transformation initiative taking place at Purdue. The IMPACT mission is to reform program courses to be engaging and student centered, thereby improving student learning, competence, and confidence. The focus on creating student-centered classrooms is derived from a deep research base on learner-centered psychological principles developed by the American Psychological Association's Board of Educational Affairs (1997). These principles synthesize the bodies of knowledge about learning and instruction, and the social and individual factors that influence the learning process. Learner-centered instructional practices are characterized by (a) the inclusion of "learners in decisions about how and why they learn and how that learning is assessed"; (b) valuing of "each learner's unique perspectives"; (c) respecting and accommodating "individual differences in learners' backgrounds, interests, abilities, and experiences"; and (d) treating "learners as co-creators and partners in the teaching and learning process" (McCombs, 2001, p. 186). These characteristics extend from a theory of learning that posits learning is an active, constructive process building upon learner prior knowledge and experience and is mediated by social interactions in the learning environment. The first and fourth characteristics described by McCombs address the fostering of student autonomy and equal partnership in the learning process so that students become motivated, engaged learners; the second and third highlight the importance of recognizing and building upon students' interests, experiences, and existing knowledge to enable them to construct their own understanding. Using the learner-centered characteristics and previous observation protocols as guides (e.g., the RTOP), we hoped to implement a single instrument that could be used to assess the best instructional practices gained through professional development and course reform across a broad range of disciplines. The exact criteria present in an "engaged student-centered" classroom varies among course transformation efforts, but most assessment measures target the degree of student engagement in lessons, discussions between students and instructors, discussion among students, and student ownership of course material and the learning process (Fraser, 2012; S. R. Harper & Quaye, 2010; Wolters & Taylor, 2012). Examples of these domains can be found on the Constructivist Learning Environment Survey developed by Taylor and Fraser (Taylor, Fraser, & Fisher, 1997). The Constructivist Learning Environment Survey demonstrates strong psychometric properties for measuring the constructivist nature of the learning environment, but the focus and testing of this protocol was limited to science and mathematics courses. The College and Univer-

sity Classroom Environment Inventory developed by Fraser focusing on the "psychosocial environment" within the classroom provides an assessment tool suited to measuring student involvement, cohesiveness, and satisfaction but was designed with the intention for use in "small higher education classes often referred to as seminars" (Fraser & Treagust, 1986, p. 37).

During our review of the existing measurement protocols, it became clear that no single instrument was suited, both in terms of content measured and design procedures employed, for addressing the broad nature of the reform initiative we were attempting with IMPACT. Themes present on existing protocols (i.e., student engagement in lessons, discussions between students and instructors, discussion among students, and student ownership of course material and the learning process) align with the fundamental goals for IMPACT as well as Purdue's experience developing reformed educational practices in STEM. However, from the outset, the scope of the IMPACT project necessitated an extension of the knowledge and experience gained from the extant reform strategies to a larger campus community. We incorporated the strengths from various instruments, focusing on the common elements present in each. The culmination of our efforts resulted in the development of two research protocols. The first took shape as an observational instrument. Simultaneously, the adopted concepts were formatted into a student self-report survey.

Our survey protocol is named the Classroom Experience Questionnaire (CEQ). The CEQ was catalyzed by the instruments named; however, the CEQ was specifically designed to address the temporal and fiscal limitations of observation research as well as the limitations of instructor self-reporting methods (Hill et al., 2012). The process of creating the CEQ involved an adaptation and transformation of constructs present on the cited protocols, such that selected items were translated into classroom practices that could be measured by student self-report as well as through observational methods. One of the primary differences between the CEQ and established instruments was making the CEQ more easily interpretable by college students reporting via a web-based survey. We developed descriptors of classroom practices that could be recorded by observers, descriptors that were indicative of a learner-centered classroom, but these descriptors were also designed to be independent of course discipline. The descriptors were written in such a way that they could also be rated by students. These descriptors primarily dealt with instructor practice or class management relating to student-centered, community-based, and engagement-oriented teaching and learning. Items were reviewed for face validity by instructors and students, revised and reconfigured.

We hoped that the CEQ would reflect the strengths of established, valid, and reliable assessment instruments while increasing the applicability of the CEQ instrument to a wider variety of academic disciplines. Further, we hoped that the modifications would allow us to rely on student perception data through a self-report survey methodology to measure the efficacy of reform and avoid the costs associated with administering an observation protocol.

The CEQ was developed around three domains that are reflective of the learner-centered characteristics described: autonomy-supporting learning climate (hereafter termed Learning Community), active construction of knowledge that builds on student prior understanding and experience (hereafter termed Constructivist Pedagogy), and equitable sharing of power and responsibility in the classroom (hereafter termed Equity).

Learning Community (LC) is measured using six items reflecting the students' perceptions of the student-centered orientation of classroom processes. Items are Likert scaled from 5 (*strongly*

agree) to 1 (*strongly disagree*) with 3 (*not sure/undecided*) as the centering option. All six items begin with the phrase "The instructor" and are as follows:

1. Encouraged students to learn from one another.
2. Provided opportunities for student to challenge opinions expressed in class.
3. Encouraged student to participate actively in class.
4. Provided opportunities for students to ask questions.
5. Allowed students to answer a question or solve a problem in more than one way.
6. Maintained a climate of respect within the class for what others had to say.

Constructivist Pedagogy (CP) reflects active construction of knowledge that builds on student prior understanding and experience. Our adaptation of this popular theme consists of four items. The four items that make up this construct are also Likert scaled from 5 (*strongly agree*) to 1 (*strongly disagree*) with 3 (*not sure/undecided*) as the centering option. Each item again begins with the phrase "The instructor" and reads as follows:

1. Connected course content to students' experience and knowledge.
2. Asked students to explain their ideas.
3. Gave students adequate time to think about and/or discuss a new concept.
4. Provided opportunities for students to process new information.

The final construct of the CEQ, Equity (EQ), focuses on student perceptions related to sharing of power in the classroom. This construct consists of three items measured using a 10-point scale with 10 representing completely instructor centered and 1 representing completely learner centered. The anchors for this scale are 10 (*instructor*), 5 (*both equally*), 1 (*students*). The items of this construct are as follows:

1. During the past week, who primarily guided the class discussion?
2. Discussion in class generally followed which format?
3. During the past week, who primarily determined the topics covered during class?

RESEARCH QUESTIONS

Of primary interest is the viability of the CEQ instrument as a tool for assessing the enactment of a learner-centered classroom. The following research questions and hypotheses framed our analysis:

1. Does the CEQ measure the intended characteristics of learner-centered instruction?
   H1: The CEQ has face validity and reflects three primary characteristics of learner-centered instruction with three main construct areas inspired by established research protocols. When run through a dimension reduction, factor analysis, and Cronbach's alpha scaling CEQ data will reproduce theorized constructs: a community learning construct, a constructivist pedagogy construct, and a classroom equity construct.

2. Do observers rate reformed courses under review in the same way?
   H2: Based on careful calibration observers will rate courses in an equivalent manner, demonstrated by statistically significant interrater reliability.
3. Are the student self-reported data a viable alternative to observer evaluations?
   H3: Observation and self-reported CEQ data will measure the student environment in a statistically equivalent manner, demonstrating concurrent validity between the observation data and the student self-reported CEQ data.

DATA AND METHODS

During the first and second wave of data collection, the IMPACT assessment team used trained observers to collect data using the CEQ measures. Observers were recruited from staff members of the university units involved with the IMPACT project. Training generally followed practices recommended for use with existing observation protocols (e.g., Sawada et al., 2000). Observers were calibrated using prerecorded instruction including training videos for observer calibration. Individual observational rating with the CEQ instrument was then conducted, followed by in-depth discussion of each rating item until a consensus was reached. Both observation and self-report data were then collected in IMPACT classrooms. The goal for collecting both data types in the same class settings was to establish concurrent validity between the student self-report CEQ and the trained observer CEQ. Observers attended and reported on one class each week. Student self-reported data were also collected each week from a random sample of the course roster. The observer data and the self-report data were then compared to determine if the self-report CEQ was a viable alternative to observation collection methods.

Observations were conducted by 13 calibrated observers. The first round of observations lasted approximately 3 months beginning in September 2011 and ending the last week of November 2011. During this first round of data collection, seven courses were from the colleges of science, liberal arts, health and human sciences, engineering, and technology, with 24 separate sections participating. In total there were 884 completed student self-report surveys and 72 unique classroom observations. Table 1 presents the makeup of the courses included in the analysis reported in this article.

ANALYTIC STRATEGY

Following the first phase of data collection the analysis plan proceeded in three steps. First, the student self-report data were tested to establish the CEQ as a valid and reliable measure of the three intended constructs. Second, the observation data were tested for interrater reliability (IRR) between observers (Zegers, 1991). Third, concurrent validity between the observation data and the student self-report data was run using the observational methods as the "gold standard" and the CEQ student self-report as the comparison.

Scales generated by CEQ data were developed through a process of dimension reduction consisting of a principle components analysis (PCA) and confirmatory factor analysis (CFA)

TABLE 1
Descriptive Statistics for Courses Included in Analysis, Numbers Reported Are Self-Report Totals

Department	*Course*	*Sections in Study*	*Total Enrollment*[a]	*N of Participations*	*Control Class*[b]
Agronomy	Soil Science (AGRY 255)	2	304	83	No
Agronomy	Genetics (AGRY 320)	1	141	121	No
Chemistry	General Chemistry (CHM 115)	2	923	285	Yes
Communication	Principles of Persuasion (COM 318)	1	344	103	No
Math	Algebra and Trigonometry II (MA 154)	1	163	57	Yes
Political Science	Intro to Political Science (POL 101)	1	170	74	Yes
Psychology	Elementary Psychology (PSY 120)	16	712[c]	161	No
		24	2,757	884 (32%)	3

[a]Including all sections. [b]Included an experimental control class that was not reformed and will be used in future analysis to report on the efficacy of redesign on our campus. [c]16 recitation sections of PSY 120 participated.

consisting of principle axis factoring (PAF) with Promax rotation accounting for correlation among constructs. PCA was first run to account for all of the variance represented by the 13 measured items, freely allowing constructs to emerge. PAF was next run to account for the shared variance within the limited substantive constructs (i.e., LC, CP, and EQ). A measure of Cronbach's alpha was also included as a measure of internal reliability.

To determine IRR a two-way mixed model of all observers was run. A two-way mixed model was necessary because the study design called for data collection from a fixed group of observers and a random sample of students (Haber, Barnhart, Song, & Gruden, 2005; Hill et al., 2012; Shrout & Fleiss, 1979).[2] Based on the scaled nature of the Likert items reliability testing was done using a series of intraclass correlation coefficients (ICC) to measure IRR.[3] Based on the categorical-Likert nature of the scaled items Goodman and Kruskal's gamma coefficient gets presented first. Spearman's correlation coefficients as well as Pearson's R are also included since measures are combined in scales. ICC ranges from 0.0 to 1.0, with higher values indicating smaller variance between observers and more agreement of raters (Hallgren, 2012).

Concurrence between the observation data and self-reported data was also measured using Goodman-Kruskal's gamma, and the Spearman, and Pearson *R* correlation coefficients. Based on the continuous nature of scaled Likert items the Spearman and Pearson correlations were considered most useful. To further increase the robustness of results Pearson's correlations were based on Bartlett factor regression scores created during PAF (for more on scale scores,

[2]Based on the IMPACT assessment design, the observer correlations cannot be interpreted as generalizable beyond these data.

[3]Another option would have been to use a weighted Kappa to measure IRR. However, ICC was relied upon because ICC takes into account the differences in individual ratings while accounting for the correlation between raters. In addition, the ICC is robust to clustering of error due to measurement by the same individuals across time varying class settings. For more on the strengths of ICC when dealing with longitudinal or mixed designs see pages 200 to 210 in Part 3 of Gwet (2012).

see Wu, 2007). The correlation pattern and range of the gamma, Spearman, and in particular the Pearson coefficients were used to determine concurrence.

RESULTS

Test of Hypothesis 1

Step 1 began with a PCA of the CEQ data. The PCA was accomplished using an oblique rotation to allow factors to be correlated. During PCA all 13 items were run in a component matrix. The expected three-factor solution emerged, though as Table 2 indicates there was considerable overlap among the items. However, the results of the PCA show that these data form around the three main factor structures representative of the intended conceptual constructs. This supports our first hypothesis.

Following the PCA, a PAF was run for each of the substantive constructs (i.e., LC, CP, and EQ). The first CFA was run on the six observed items that make up the LC construct. Table 3 presents these results. PAF results show that these items form around a strong construct. For

TABLE 2
Principle Components Analysis With Standardized Factor Loading of Classroom Experience Questionnaire Data

	Observation Data[a]			*Self-Reported Data*[b]		
Component	*1*	*2*	*3*	*1*	*2*	*3*
Item 1	.620	.394	.385	−.116	.740	
Item 2	.510	.702	.140	.729		
Item 3	.676	−.386	.410	.852		
Item 4	.721	−.325		.854	−.193	
Item 5	.627	−.552	−.281	.768		
Item 6	.848	.126		.811		
Item 7	.751	.178		.817		
Item 8	.603	−.483		.715		.169
Item 9	.851		−.199		.789	
Item 10	.343	.182	.603	.695	−.178	−.251
Item 11	.742	.231	−.363	.587	.105	.231
Item 12		−.333	.413			.946
Item 13	.502	.195	−.559	.619	.210	.103
	Promax Rotation Component Correlation Matrix					
Component	*1*	*2*	*3*	*1*	*2*	*3*
1	1.000			1.000		
2	.453	1.000		.321	1.000	
3	.372	.272	1.000	−.033	.047	1.000

Note. KMO = Kaiser-Meyer-Olkin.
[a]$n = 71$; KMO and Bartlett Test Statistic = .684. [b]$n = 649$; KMO and Bartlett Test Statistic = .903. KMO and Bartlett = $p \leq .001$.

TABLE 3
Principle Axis Factor With Standardized Factor Loading of the Learning Community Construct

	Observation Data[a]	*Self-Reported Data*[b]
KMO and Bartlett Test statistic =	.804***	.833***
The Instructor:	*Factor Loadings*	*Factor Loadings*
Item 1: Encouraged students to learn from one another.	.772	.499
Item 2: Provided opportunities for students to challenge opinions expressed in class.	.618	.622
Item 3: Encouraged students to participate actively in class.	.893	.799
Item 4: Provided opportunities for students to ask questions.	.697	.747
Item 5: Allowed students to answer a question or solve a problem in more than one way.	.778	.705
Item 6: Maintained a climate of respect within the class for what others had to say.	.319	.592
Variance explained	49.48%	44.63%
Cronbach's α	.843	.820

Note. n = six items factored. KMO = Kaiser-Meyer-Olkin.
[a]$n = 71$. [b]$n = 649$.
***$p \leq .001$.

instance, PAF using the observation data returned a Kaiser-Meyer-Olkin (KMO) and Bartlett Test[4] statistic of .804 ($p = .001$) with loadings ranging from .319 low to .893 high, 50% of the variance explained, and a Cronbach's alpha of .843. Student data returned a KMO and Bartlett Test statistic of .833 ($p = .001$) with loadings ranging from .499 low to .799 high, 47% of the variance explained, and a Cronbach's alpha of .820.

Next, Table 4 displays the results of a CFA that was run on four items reflecting active construction of knowledge building on student prior understanding and experience. This is the CP construct. PAF using the observation data returned a KMO and Bartlett Test statistic of .624 ($p = .001$) with loadings ranging from .392 low to .985 high, 56% of the variance explained, and a Cronbach's alpha of .798. Student data returned a KMO and Bartlett Test statistic of .759 ($p = .001$) with loadings ranging from .664 low to .852 high, 58% of the variance explained, and a Cronbach's alpha of .839.

The final construct of the CEQ measures equitable sharing of power and responsibility in the classroom. This is the EQ construct. PAF using the observation data returned a KMO and Bartlett Test statistic of .616 ($p = .001$) with loadings ranging from .430 low to .738 high, 40% of the variance explained, and a Cronbach's alpha of .595. Student data returned a KMO and Bartlett Test statistic of .556 ($p = .001$) with loadings ranging from .405 low to .987 high, 46% of the variance explained, and a Cronbach's alpha of .630. CFA results of EQ returned a moderate (at best) construct in these data. Table 5 displays the results of this CFA and provides additional information regarding missing values.

[4]The Kaiser-Meyer-Olkin measure of sampling adequacy tests the partial correlations among factor variables. Bartlett's test is a measure of sphericity testing whether the correlation matrix is an identity matrix.

TABLE 4
Principle Axis Factor With Standardized Factor Loading of the Constructivist Pedagogy (CP) Construct

	Observation Data[a]	*Self-Reported Data*[b]
The Instructor:	*Factor Loadings*	*Factor Loadings*
Item 7: Connected course content to students' experience and knowledge.	.392	.693
Item 8: Asked students to explain their ideas.	.610	.664
Item 9: Gave students adequate time to think about and/or discuss a new concept.	.985	.809
Item 10: Provided opportunities for students to process new information.	.859	.852
Variance explained =	55.48%	57.54%
Cronbach's α =	.798	.839

Note. n = four items factored. KMO = Kaiser-Meyer-Olkin.
[a]n = 71; KMO and Bartlett Test statistic = .624. [b]n = 649; KMO and Bartlett Test statistic = .759. KMO and Bartlett = $p \leq .001$.

Table 5 shows that for the EQ construct, missing data points are problematic. A skip-pattern in the design largely accounts for the holes in these data. Prior to seeing the three questions that compose the EQ construct, observers and students are presented with the following question, "There was a discussion portion in class this week (1 = Yes 2 = No)?" Respondents who answer this question in the negative are skipped past the first question (Item 11) on the EQ scale. According to this skip-pattern we see that 62% of observers and 45% of students perceived that

TABLE 5
Principle Axis Factor With Standardized Factor Loading of the Equity Construct

	Observation Data[a]	*Self-Reported Data*[b]
	Factor Loadings	*Factor Loadings*
Item 11: During the past week, who primarily guided the class discussion?	.738	.487
Item 12: Discussion in class generally followed which format (format options given)?	.674	.987
Item 13: During the past week, who primarily determined the topics covered during class?	.430	.405
Variance explained	39.46%	45.86%
Cronbach's α	.595	.630
Item 11% missing cases	62.0%	44.74%
Item 12% missing cases	59.2%	35.41%
Item 13% missing cases	2.8%	25.33%

Note. n = three items factored. KMO = Kaiser-Meyer-Olkin.
[a]n = 71; KMO and Bartlett Test statistic = .616. [b]n = 649; KMO and Bartlett Test statistic = .556. $p \leq .01$; KMO and Bartlett = $p \leq .001$.

no discussion happened during class. In terms of reform assessment, this is a noteworthy finding by itself (Brown, Furtak, Timms, Nagashima, & Wilson, 2010; Clare & Aschbacher, 2001). As a result of this skip-pattern the KMO/Bartlett statistic indicated that sampling was not adequate for this construct. Items 12 and 13 in Table 5 are a part of another skip-pattern occurring at a later point in the CEQ survey. Missing data related to the measurement of EQ made it impossible to accurately analyze this construct using these data. Despite the disappointing result of the EQ construct, largely caused by the missing data points, we found support for our initial hypothesis: Each proposed construct emerged as a good measure of the intended domains. Next we used the CEQ measures to test the IRR of the observation data.

Test of Hypothesis 2

Step 2 was our test of IRR among observers. Our test of IRR returned an intraclass correlation coefficient (ICC) for single measures of .726 ($df = 8$, $p = .001$) and for averaged measures an ICC of .995 ($df = 8$, $p = .001$). Cronbach's scale of the IRR was .998 ($n = 71$ observations). These results strongly support our hypothesis of IRR in these data. Next, using the CFA results from our observer data, we used the scaled values from the LC construct as a specific measure of IRR between observers. Using individual observer measurement of the LC scale confirmed the single measure ICC of .726 that was initially significant ($p = .001$) in our test of IRR. These results strongly indicate that IRR was achieved among the observers, confirming our second hypothesis.

Test of Hypothesis 3

The final step of the analysis involved a comparison of observation data with the self-reported data to establish concurrent validity. Observation data achieved IRR, and the student self-reported data created the same constructs to those found in the observation data. Having achieved these encouraging results, the third and final step of our analysis was to test concurrence between the observer ratings and the students' ratings. To measure concurrence between the observation data and self-reported data multiple dependence coefficients were computed. Dependence between the Likert summative scales in each data set was measured using gamma, Spearman, and Pearson correlation coefficients. Strong correlation, as shown in Tables 6 and 7, was found between the observation data and the self-reported data ($\gamma = .597–.653$, $p \leq .001$; Spearman $= .830–.792$, $p \leq .001$; Pearson $= .826–.788$, $p \leq .001$). Recall, coefficients

TABLE 6
Concurrent Validity Statistics for Observation Data

		Value	*Asymptotic SE*	*Approx. T*
Ordinal by ordinal	Gamma	.597***	.047	12.714
	Spearman correlation	.792***	.046	10.466
Interval by interval	Pearson's *R*	.788***	.043	10.325
	N of valid cases	67		

***$p \leq .001$.

TABLE 7
Concurrent Validity Statistics for Self-Reported Data

		Value	*Asymptotic SE*	*Approx. T*
Ordinal by ordinal	Gamma	.653***	.016	42.757
	Spearman correlation	.830***	.014	37.378
Interval by interval	Pearson's *R*	.826***	.013	36.792
	N of valid cases	632		

***$p \leq .001$.

range from 0 to 1, with results approaching 1 representing more agreement between ratings. The consistently strong result across correlation measures demonstrates the robustness of our findings. These results support concurrent validity between the observation and self-reported data generated by the CEQ, supporting our third and final hypothesis. After careful instrument creation, calibration of observers, and randomized student participation, we found that observers and students independently measured the learning environment in a statistically equivalent manner. This is a noteworthy result.

DISCUSSION

The results of our analyses suggest that the CEQ possesses strong construct validity, internal reliability including interrate reliability, and internal consistency, and has concurrent validity between observation data and student self-reported data.

The overall results are a good indication that the CEQ effectively reproduces constructs that frequently appear on assessment protocols designed to measure the student-centeredness of a higher education learning environment. Based on our review of extant protocols we believe the CEQ possess face validity and that our findings support this conclusion; the CEQ also has content and construct validity. The two constructs with sufficient sampling, *Learning Community* and *Constructivist Pedagogy*, possessed strong internal consistency as evidenced though PAF loadings, variance explained, and Cronbach's alpha. Based on these results, we advance the CEQ as a valid measure of learner-centered constructivist instruction and that the weight given to these two student self-report constructs should be treated as equivalent to the weight given to the same constructs measured through observation methods.

There is a significant amount of research demonstrating that students evaluate classroom practices differently than professional observers (Centra, 1993; Isaacson et al., 1964; Simpson & Siguaw, 2000; Sojka, Gupta, & Deeter-Schmelz, 2002). We believe our findings suggest that student self-reported data are a valid alternative for rating the level of engagement and student-centeredness of a course but do not argue for equivalence beyond the measured domains (Schunk & Meece, 1992). Determining the utility of student self-report data as an alternative to observation methods was a primary motivational factor behind the creation of the CEQ. However, the CEQ specifically focuses on domains relating to student-centeredness and engagement. This necessarily limits the scope and utility of the CEQ, and in many ways the CEQ is narrower than many established protocols. Further, our results are not meant to

suggest that students and observers will always be able to assess instructional practices in an equivalent way. What we are suggesting is that when it comes to evaluating the degree to which a course engages students with the content and community of learners in the classroom, students are a reliable source of measuring this information. Concurrence between observer CEQ data and student self-reported data is an important result. In terms of the LC and CP constructs, students rate this aspect of the learning environment in a statistically equivalent manner to trained observers.

Use of the CEQ is not limited to reformed courses. CEQ measures relate to the content and community of learners including constructivist pedagogy as well as the student-centered nature of a course. The CEQ could be used as a measure of these domains in reformed and nonreformed courses alike. We saw a need for a self-report instrument that was a reliable measure of reformed teaching and learning practices in the university classroom, an instrument not tied to discipline-based approaches to teaching and learning. The CEQ was validated across a broad range of disciples, and we believe our results indicate that the CEQ can be used to measure the extent to which instructor practice employs the hallmarks of learner-centered instructional reform in a broad range of courses. Blended, learner-centered approaches to instruction are becoming more common in higher education and, as is the case at our university, are being applied to very large, foundational courses across disciplines. It is important to continue to develop measurement methods that allow for increased flexibility due to the costs and logistical difficulties of employing classroom observations in these settings.

Although a constructivist view of learning is accepted across disciplines, it has historically been more widely advanced and studied in the science and mathematics disciplines. The development of the CEQ extends assessment of constructivist practices beyond these disciplines so that course reform can be assessed across an institution. This is particularly important, as course reform efforts are executed at the university level, rather than the college, school, or department level.

LIMITATIONS AND SUGGESTIONS

The missing data results reported raise a couple of additional questions that need to be addressed moving forward. For instance, (a) with nearly 50% of courses reporting, no discussion how student centered are these classes, and (b) can the CEQ measures of engagement and student-centered learning be used to predict student learning gains? In reference to the first question, there is a distinct possibility that students may have misinterpreted the meaning of “discussion” in the CEQ items. In a traditional context, discussion would occur within the physical classroom and would have been readily observed by both students and trained observers. If discussions were occurring outside of the classroom, using a technological medium such as a discussion board within a learning management system, then trained observers would not have indicated in their evaluations that these discussions occurred. It is also possible that the students did not perceive online discussions to be the same as “class discussions” as referenced in the CEQ items. Future work on the CEQ, as well as other course assessment tools, will need to address the complexities of modern reform practices that move pedagogy to online spaces that are not readily apparent when assessing class time. These ideas and these findings may result in a revision to the CEQ items for improved measurement purposes relative to classroom

"discussion" sections of the CEQ as well as other evolving areas of the contemporary higher education classroom. Despite this limitation, the CEQ was created to measure engagement and student-centered learning in the classroom environment; at present the instrument is sensitive to this level of measurement, but based on its current construction is limited to this environment.

In reference to the second question, can the CEQ measures of engagement and student-centered learning be used to predict student learning gains? As our data continue to be gathered, future analysis efforts will, among other things, use the CEQ as a predictor of student achievement. Linking student self-reports of the learning environment to student achievement is the next important step for determining the utility of the CEQ and is an area requiring further study.

Data collection for the IMPACT project is ongoing and includes classical experimental methods where students in IMPACT courses are compared to students in control courses. Data currently being gathered also include longitudinal measures of CEQ constructs as well as student achievement measures such as course grades and retention rates. When IMPACT began, one of our initial tasks was finding or establishing a survey protocol that could be used to measure student centeredness and student engagement for a broad-sweeping redesign initiative. The process of validating the CEQ as a measure of the learning climate as well as determining if trained observers were statistically equivalent to student self-reports was the main focus of this article. Our next steps will use the CEQ measures along with demographic and other control variables to predict student outcomes such as grades, drop, fail, and withdrawal rates, as well as graduation and placement. This future research will make use of the classical experimental design scenarios in place as well as longitudinal measurement of courses, comparing the efficacy of redesign from prereform to postreform.

FUNDING

This research was supported in part by funding from the Office of the Provost, Purdue University.

REFERENCES

Arum, R., & Roksa, J. (2010). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Astin, A. W. (1971). Two approaches to measuring students' perceptions of their college environment. *Journal of College Student Personnel, 12*, 169–172.

Astin, A. W. (1977). *Four critical years. Effects of college on beliefs, attitudes, and knowledge*. San Francisco, CA. Wiley, Jossey-Bass.

Astin, A. W., & Astin, H. S. (1992). *Undergraduate Science Education: The Impact of Different College Environments on the Educational Pipeline in the Sciences*. Final report.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62–87.

Blackie, M. A. L., Case, J. M., & Jawitz, J. (2010). Student-centredness: The link between transforming students and transforming ourselves. *Teaching in Higher Education, 15*, 637–646.

Blickenstaff, J. C. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education, 17*, 369–386.

Borrego, M., Streveler, R. A., Miller, R. L., & Smith, K. A. (2008). A new paradigm for a new field: Communicating representations of engineering education research. *Journal of Engineering Education, 97*, 147–162.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The evidence-based reasoning framework: Assessing scientific reasoning. *Educational Assessment, 15*, 123–141.

Carnell, E. (2007). Conceptions of effective teaching in higher education: extending the boundaries. *Teaching in Higher Education, 12*, 25–40.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. The Jossey-Bass higher and adult education series*: ERIC.

Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*, 39–59.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*, 113–143.

De Welde, K., Laursen, S., & Thiry, H. (2007). Women in science, technology, engineering and math (STEM). *ADVANCE Library Collection*.

Dean, D. J., & Fleckenstein, A. (2007). Keys to success for women in science. In R. J. Burke & M. C. Mattis (Eds.), *Women and minorities in science, technology, engineering and mathematics: Upping the numbers* (pp. 28–44). Northampton, MA: Edward Elgar.

Denton, D. D. (1998). Engineering education for the 21st century: Challenges and opportunities. *Journal of Engineering Education, 87*, 19–22.

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

Fraser, B. J. (2012). Classroom learning environments: Retrospect, context and prospect. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (Vol. 24, pp. 1191–1239). Springer Netherlands.

Fraser, B. J., & Treagust, D. F. (1986). Validity and use of an instrument for assessing classroom psychosocial environment in higher education. *Higher Education, 15*, 37–57.

Freire, P. (2000). *Pedagogy of the oppressed: 30th anniversary edition*: Bloomsbury Academic.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. National Bureau of Economic Research.

Gwet, K. L. (2012). *Handbook of inter-rater reliability* (*3rd edition*): *The definitive guide to measuring the extent of agreement among multiple raters*. Gaitherburg, MD: Advanced Analytics, LLC.

Haber, M., Barnhart, H. X., Song, J., & Gruden, J. (2005). Observer variability: A new approach in evaluating interobserver agreement. *Journal of Data Science, 3*, 69–83.

Haghighi, K. (2005). Systematic and sustainable reform in engineering education. *Journal of Environmental Engineering, 131*, 501–502.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol, 8*, 23–34.

Harper, S. R., & Quaye, S. J. (2010). *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*. New York, NY: Taylor & Francis.

Harper, V. B., Jr. (2009). Virginia's value added: A diverse system perspective. *Assessment Update, 21*(4), 1–2.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Lynch, K., . . . (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*, 88–106.

Hughes, G. (2007). Using blended learning to increase learner support and improve retention. *Teaching in Higher Education, 12*, 349–363.

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., & Zinn, K. L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology, 55*, 344.

Jenny, H. H. (1996). *Cost accounting in higher education. Simplified macro-and micro-costing techniques.* ERIC.

Katehi, L., Banks, K., Diefes-Dux, H., Follman, D., Gaunt, J., Haghighi, K., . . . Oakes, W. (2004). *A new framework for academic reform in engineering education.* Paper presented at the American Society for Engineering Education Conference, Salt Lake City, UT.

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology, 14*, 366–381.

King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal, 27*, 664–687.

King, A. (1993). From sage on the stage to guide on the side. *College Teaching, 41*, 30–35.

Kuh, G. D. (2001). *The national survey of student engagement: Conceptual framework and overview of psychometric properties* 1–26. Bloomington: Indiana University Center for Postsecondary Research.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning, 35*(2), 24–32.

McCombs, B. L. (2001). What do we know about learners and learning? The learner-centered framework: Bringing the educational system into balance. *Educational Horizons, 79*, 182–193.

McCray, R. A., DeHaan, R. L., & Schuck, J. A. (Eds.). (2003). *Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop.* Washington, DC: National Academies Press.

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment, 16*, 227–243.

Miller, J. E., & Groccia, J. E. (2011). *To improve the academy: Resources for faculty, instructional, and organizational development.* Wiley.

The National Center for Academic Transformation: Who We Are. (2013). Retrieved from http://www.thencat.org/whoweare.html

NSF. (2012). Science and Engineering Indicators 2012. From http://www.nsf.gov/statistics/seind12/c2/c2s2.htm

Office of Management and Budget. (2013). *Fiscal Year 2014: Budget of the U.S. Government*. Washington, DC: Government Printing Office.

Pace, C. R. (1984). *Measuring the quality of college student experiences. An account of the development and use of the College Student Experiences Questionnaire*.

Pace, C. R. (1985). *The credibility of student self-reports*.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP) reference manual.* Tempe, Arizona: Arizona Collaborative for Excellence in the Preparation of Teachers.

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. *New Directions for Institutional Research, 2011*(150), 41–58.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement, 53*, 801–813.

Potočnik, J. (2009). Women in science and technology: Creating sustainable careers In O. f. O. P. o. t. E. Communities (Ed.). Luxembourg, Belgium: European Communities.

Provezis, S. (2010). Regional accreditation and student learning outcomes: Mapping the territory. *National Institute for Learning Outcomes Assessment, Occasional Paper*(6), 7.

Sawada, D., Piburn, M., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP) training guide* (ACEPT IN-002). Arizona Board of Regents. Retrieved from https://mathed.asu.edu/instruments/rtop/Training_Guide_Mar2000.pdf

Sawada, D., Piburn, M. D., & Judson, E. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science & Mathematics, 102*, 245–253.

Schunk, D. H., & Meece, J. L. (1992). *Student perceptions in the classroom.* Hillsdale, NJ: Erlbaum.

Shrout, P. E, & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Simpson, P. M., & Siguaw, J.A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*, 199–213.

Slavin, Robert E. (1995). *Cooperative learning: Theory, research, and practice* (Vol. 2). Boston, MA: Allyn and Bacon.

Smith, M. K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices. *CBE-Life Sciences Education, 12*, 618–627.

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching, 50*, 44–49.

Streveler, R. A., Litzinger, T. A., Miller, R. L., & Steif, P. S. (2008). Learning conceptual knowledge in the engineering sciences: Overview and future research directions. *Journal of Engineering Education, 97*, 279–294.

Swing, R. L., & Coogan, C. S. (2010). *Valuing assessment: Cost-benefit considerations*. Champaign, IL: National Institute for Learning Outcomes Assessment.

Taylor, P. C., Fraser, B. J., & Fisher, D. L. (1997). Monitoring constructivist classroom learning environements. *International Journal of Educational Research, 27*, 293–302.

Turner, P. M., & Carriveau, R. S. (2010). *Next generation course redesign.* Peter Lang.

Turner, P. M. (2009). Next generation: Course redesign. *Change: The Magazine of Higher Learning, 41*(6), 10–16.

Twigg, C. A. (2006). *Improving learning and reducing costs: Project outcomes and lesson learned from the roadmap to redesign*. Program in Course Redesign: Round I: The National Center for Academic Transformation.

Watson, K., & Froyd, J. (2007). Diversifying the US engineering workforce: A new model. *Journal of Engineering Education, 96*, 19–32.

Wolters, C. A., & Taylor, D. J. (2012). A self-regulated learning perspective on student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 635–651). New York, NY: Springer.

Wu, C.-H. (2007). An empirical study on the transformation of Likert-scale data to numerical scores. *Applied Mathematical Sciences, 1*, 2851–2862.

Zegers, F. E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement, 15*, 321–333.

Zimpher, N. L. (2009). The leaky pipeline: IT can help. *EDUCAUSE Review, 3*, 4–5.