Measuring Students' Perception of Learning: The Systematic Development of an Instrument

Daniel G. Ngugi, Lisa Borden-King, Draza Markovic, Andy Bertsch

Minot State University

Within the education sector various tools have been used to measure effectiveness of instruction. It is typical that measures of teaching effectiveness include, but are not limited to, the student's perception of their experience in the classroom and with a given instructor. Student evaluations of teaching (SETs) are one form of measurement commonly used in American universities. It is important to determine whether these SETs are helpful in assessing effective teaching and the instructor's work in and out of the classroom, in general. To determine whether these SETs are helpful in assessing effective teaching and the instructor's work in and out of the classroom, in general, we sought to develop an instrument to measure the students' perception of teaching and learning as represented by three concepts: Student, Course, and Instructor. We used scaled survey items, some of which we borrowed from other instruments to operationalize the concepts and create a pilot test. We analyzed the data using Factor analysis techniques. The result was an instrument that included 24 items scaled on a five-point Likert scale.

Dans le secteur de l'éducation, divers outils ont servi à l'évaluation de l'efficacité de l'enseignement. Typiquement, les mesures de l'efficacité de l'enseignement comprennent, entre autres, la perception qu'a l'étudiant de son expérience en classe et avec son professeur. Les évaluations par les étudiants de l'enseignement sont une mesure couramment utilisée dans les universités américaines. Il est important de déterminer si ces évaluations par les étudiants sont utiles dans l'évaluation générale de l'efficacité de l'enseignement et du travail du professeur en salle de classe et à l'extérieur de celle-ci. Pour le faire, nous avons tenté de développer un instrument permettant de mesurer la perception qu'ont les étudiants de l'enseignement et de l'apprentissage en fonction de trois concepts : l'étudiant, le cours et le professeur. Pour mettre en œuvre les concepts et créer un essai pilote, nous nous sommes servis de questions de sondage échelonnées, dont certaines ont été empruntées à d'autres instruments. Nous avons analysé les données avec des techniques d'analyse factorielle. Le résultat est un instrument à 24 items gradués selon l'échelle de Likert.

Much has been written concerning the relationship between students' perceptions of learning, their approaches to learning, their ambition to learn, their motivation to learn, the context in which that learning takes place, and other variables in the relationship between teaching and learning (see, e.g., Bertsch et al., 2016; Entwistle, McCune, & Hounsell, 2003; Entwistle & Ramsden, 1983; Marton & Säljö, 1976, 1984). There is a robust research stream dedicated to the relationship between “surface” versus “deep” learning, students’ perception of learning, the learning context, students’ perception of good teaching, and actual good teaching (Parpala, Lindblom-Ylänne, Komulainen, & Entwistle, 2013). Likewise, there is research exploring intrinsic and extrinsic academic motivation (Bertsch et al., 2016). This study seeks to explore a succinct instrument to be used as part of faculty annual evaluation performed by university administration.

The purpose of this study was limited in scope and focused on reviewing student evaluation of teaching instruments (SETs). The primary outcome of this effort was to develop a succinct instrument to measure the students’ perception of teaching and learning as represented by three specific concepts: (i) the student concept—the student’s perception of their own preparedness, engagement, and commitment to their own learning; (ii) the course concept—the course content, course organization, breadth of material covered, and the overall learning experience; and (iii) the instructor concept—instructor’s knowledge of the course, rapport with the students, enthusiasm, and classroom behavior. To that end, and as will be discussed in the section Conclusions, we acknowledge that additional theoretical constructs that are more robust and multidimensional have been shown to exist such as those represented on the Experiences of Teaching and Learning Questionnaire (ETLQ) (Entwistle et al., 2003).

The subject of student evaluation of teaching (SET) has generated abundant debate in education (Berk, 2014; Feldman, 1997; Marsh & Roche, 1997; Wachtel, 1998). High stakes decisions including merit reviews, tenure, and promotions often take these evaluations into consideration (Abrami, d’Apollonia, & Cohen, 1990; Watkins, 1994; Wines & Lau, 2006). To be able to understand SET, one needs to answer the question: what is effective teaching? Effective teaching may be defined simply as activities that promote student learning; it covers all instructor actions that support learning and achievement of the educational objectives (University of California, Los Angeles Office of Instructional Development [UCLAOID], 2011).

Many scholars agree that teaching is a complex activity consisting of multiple dimensions and thus evaluating it requires more than a single approach (Berk, 2014; Cathcart, Greer, & Neale, 2014; Marsh & Roche, 1997; Simmons, 1997; UCLAOID, 2011). Because there is no direct way of measuring effective teaching, the more, and varied the data sources, the more useful the assessment is likely to be. Among others, measurement approaches may include student ratings, self-reviews, peer evaluations, expert ratings review, exit and alumni ratings, and teaching scholarship (Berk, 2014).

Many instructors are skeptical and/or critical of the suitability or usefulness of SETs to provide evidence for teaching effectiveness, particularly when used singly (see for discussion, Ackerman, Gross, & Vigneron, 2009; Beran & Rokosh, 2009; Catano & Harvey, 2011; Cathcart et al., 2014; Nygaard & Belluigi, 2011; Rienties, 2014). Yet, when designed and applied properly, SETs serve an important purpose by providing the academic community with information on the teaching–learning process, and, if well utilized, can support improvement of instruction (Ackerman et al., 2009; Catano & Harvey, 2011; Cathcart et al., 2014; Golding & Adam, 2016; Marsh, 1982). It seems, however, that any evaluation of teaching effectiveness should be based only in part on the views of the student; the rest of the input must come from other sources. This article is based on a study designed to develop an instrument to measure students’ perception of their learning, as one tool for the evaluation of teaching. To do this, the literature was first reviewed to identify the intent and use of SETs. The second step was to identify, define, and operationalize concepts (measurable constructs) that have been shown to encompass the classroom experience from the students’ perspective.

The remainder of this article proceeds as follows: in the next section a review of the literature is presented. Thereafter the methodology and data collection procedures are discussed, followed by a description of data analysis and presentation of results.

Literature Review

Patton et al. (2014) view evaluation as "a systematic process to determine merit, worth, value or significance" (What is Evaluation? section). Evaluation of teaching cannot be separated from evaluation of learning, as the latter is the expected outcome of the former. Proper evaluation of teaching, really, needs to determine whether—and how—learning is taking place. Yet teaching and learning are invisible, and to some extent subjective concepts; they are not easily or directly measured; they are latent variables.

Measuring the quality of teaching has become a major component of measuring the success of academic programs and institutions (Bok, 1992; Ewell, 1991; Titus, 2008). SETs based on the rating of instructor and instruction by the students are probably the most utilized data source for evaluating teaching (Kulik, 2001) in most American institutions of higher learning (Kalender, 2015). Titus (2008) asserts that SETs are meant to improve teaching at both the institutional and instructor levels. At the institutional level SETs are often used as a summative tool, to support hiring, retention, and promotion decisions (Ackerman et al., 2009; Marsh, 1982, 2007; Spooren, Brockx, & Mortelmans, 2013; Titus, 2008). At the instructor level, well-designed and utilized SETs may be useful as a formative tool, to assist the instructor in improving on instruction (Abrami et al., 1990; Marsh, 2007; Nygaard & Belluigi, 2011; Spooren et al., 2013; Titus, 2008; Wachtel, 1998.

Ever since student-based course evaluations (aka, student evaluations of teaching) started being implemented on a massive scale in the 60s and 70s, concerns have been raised on the potential for the practice's detrimental effects on the quality of higher education. Apart from the easily identifiable vulnerabilities of SETs such as, gender and race stereotyping (see e.g. MacNell, Driscoll, & Hunt, 2015; Titus, 2008; Wines & Lau, 2006), the criticism of SETs has often focused on their potential to drive grade inflation. The Pennsylvania State University (PSU) faculty senate committee on undergraduate education defines grade inflation as "the increase in GPA that does not reflect improved student performance" (PSU, 2003, p. 89). The committee documents a steady growth in the percentage of the highest grades (A), following the implementation of students' rating of teaching effectiveness (SRTE), at the institution, beginning in 1987. Others who have discussed this apparent link between grade inflation and the onset of SETs include Aitken (2016), Johnson (2003), Rojstaczer (2002), and Rojstaczer and Healy (2010, 2012). Further, some scholars see a direct connection between grade inflation and SETs in general including: Ellis, Burke, Lomire, and McCormack (2003); Feldman (1976); Johnson (2003); and Rojstaczer and Healy (2012). Nevertheless, this discussion may still be unresolved. There may be a crucial intermediate step connecting SETs and rising grades: an actual, sustained, improvement in students' learning, leading to genuine, reality-based satisfaction, which is then expressed on the SETs. For further discussion, see Powell, Farrar, and Cohen (1985); Rojstaczer (2002); and Rojstaczer and Healy (2010, 2012).

A potentially more insidious effect of SETs was detected in the early years of their wide implementation but has not received much attention since. It is best expressed in the words of Zelby (1974) who based his conclusions on a careful experiment he had conducted in his classes at the University of Oklahoma, Norman. The results of Zelby's study indicated most students (as expressed on the SETs) preferred instruction methods that emphasize stereotypical information retrieval (or how to) tasks, at the expense of a deeper understanding of the material and underlying concepts. Zelby warned of a deleterious long-term effect that indiscriminate application of SETs might have on the quality of education.

Underlying every SET instrument (e.g., students survey) is the presumption that overt discrete behaviors of the instructor observable by students can distinguish between ineffective and excellent teaching (Loes, Salisbury, & Pascarella, 2015; Titus, 2008). The typical SET instrument uses a list of behaviors with the students being asked to rate each item on a scale (say, between *extremely good* and *extremely bad*). A summary and sometimes one key question from the survey are then used to gauge the quality or effectiveness of teaching.

The list of behaviors measured on the typical SET reflects the instructional model that the developer of the instrument assumes to constitute effective teaching or to foster effective learning. The typical SET seems to derive from the traditional teacher-centered model of teaching (Centra, 1993; Kolitch & Dean, 1999; Scherer, Nilsen, & Jansen, 2016; Titus, 2008). This model is likely based on the (apparently flawed) reasoning that education is simply transmission of information; it does not require the student's active contribution to the learning process (Freire, 1992; Titus, 2008). The critical-engaged (or simply critical) pedagogy model (see e.g., Kolitch & Dean, 1999; Titus, 2008), on the other hand, supposes that education requires the development of a critical perception in the student, developed through meaningful dialogue (Titus, 2008). It seems that both models have something to offer: quality education should probably have elements of transmission of the information as well as the development of minds that can critically analyze that information in the context of everyday living.

In the literature, the number of items and concepts included in the SET surveys varies. This number is typically between a handful and over 30 (see e.g., Ackerman et al., 2009; Catano & Harvey, 2011; Marsh, 1982 2007; Titus, 2008). The literature seems to suggest that items on the form should cover at least three broad concepts that embody the underlying factor that impact learning: student, instructor, and instruction/course (see e.g., Kolitch & Dean, 1999; Marsh, 1982, 2007; Titus, 2008; Wachtel, 1998). Marsh (1982) applied an instrument with 41 items and 9 dimensions of instruction. The dimensions include learning (value), instructor enthusiasm, organization (course/instruction), group interaction, (individual) instructor rapport, breadth of material covered, course grading, overall measures, and student comments. Kolitch and Dean (1999) applied an instrument with 19 items and 5 concepts/categories: organization of curriculum, instructor's classroom behavior, evaluation of student performance, relationships (between students and instructor), and, overall rating.

Titus (2008) examined a SET instrument created by the University of Washington and used by more than 50 postsecondary institutions in the US. The default survey has 31 questions including 12 core questions covering four aspects of teaching: course, course content, instructor, and overall teaching effectiveness. In addition to the overall effectiveness question, the list could be seen to encompass three constructs: course (i.e., course and course content), instructor, and student. A review of SET surveys used by a wide range of institutions supported the three concepts approach.

Based on the review of the literature we envision that both the instructor and the student have an important role to play in fostering learning; their behaviors influence learning in one way or another. Therefore, a suitable instrument for capturing the learner's perceptions should contain items that elicit responses on the behavior of the two parties to the learning process—the student and the instructor—as well as the course itself.

This study sought to measure the students' perception of (teaching and) learning as represented by the three concepts on which effective teaching/learning hinges: Student, Course, and Instructor. The student concept involves the student's perception of their own preparedness, engagement, and commitment to their own learning. The course concept involves the course content, course organization, breadth of material covered, and the overall learning experience. The instructor concept involves instructor's knowledge of the course, rapport with the students, enthusiasm, and classroom behavior. To keep with the tradition and for comparison, questions that seek an overall rating on each of these concepts were also included. Finally, space was provided for the respondent to expound on their perception as well as provide additional information that could further inform on the various aspects of learning by use of open-ended questions.

Materials and Methods

The study sought to measure the students' perception of teaching and learning as represented by three concepts of Student, Course, and Instructor. The concepts were operationalized and scaled by an institutional select committee (ISC), including the authors and student representatives. Survey items were either created by the committee or borrowed, with modification, from instruments used by several varied institutions. The result was an instrument that included several preliminary and demographic questions plus 24 items scaled on a 5-point Likert scale with choices ranging from *Strongly Agree* to *Strongly Disagree* and a *Not Applicable* option (see Appendix).

Data Collection and Analysis

After the survey instrument was designed, a pilot study (i.e., a series of interviews with students and faculty) was conducted. The researchers conducting this study represent seven different departments and three different colleges at a regional university in the Midwest, USA. Students and faculty from those departments were recruited to participate in this pilot study. For the purpose of this study, the interviews were meant strictly to assess face validity as offered by Hair, Black, Babin, Anderson, and Tatham, 2006: "[The] extent to which a set of measured variables [indicators] actually represent the theoretical latent construct …" (Hair, et al., 2006, p. 707). Although Hair et al. (2006, p. 771) suggest that face validity is based solely on the researcher's judgement, interviews of students and faculty were conducted in order to inquire whether the survey items intended to measure each construct were, in fact, consistent with the theoretical definition of each construct. Latent variables cannot be directly measured; instead, such variables are represented by survey indicators (Hair et al., 2006). During these interviews, the researchers walked the participants through the survey. Participants were asked which of the three theoretical constructs was being measured by each of the survey items. As described earlier, the three latent constructs included: (i) the student concept—the student's perception of their own preparedness, engagement, and commitment to their own learning; (ii) the course concept—the course content, course organization, breadth of material covered, and the overall learning experience; and (iii) the instructor concept—instructor's knowledge of the course, rapport with the students, enthusiasm, and classroom behavior. The complete survey is provided as an appendix herein. The results of these interviews indicated universal agreement that the survey items were representative of the theoretical constructs as presented.

After completing the pilot test, a convenient but representative sample of courses was selected and applied to further test the instrument. The intent was to sample enough respondents to test the reliability and validity of the newly formed instrument. A convenience sampling method was used to recruit faculty members to distribute the pilot evaluation form in their courses. All the students present in the recruited faculty member's course at this stage of the study were surveyed. Within exploratory research designs, the sample is often based on convenience (see, e.g., Hair, Babin, Money, & Samouel, 2003; Malhotra, 2007; Zikmund & Babin, 2007). The pilot SET form was tested across 11 departments/disciplines with at least three courses at each (100, 200, 300, and 400) level. In addition to completing the forms, students as well as faculty were also asked to provide feedback on the form and its potential for capturing the three concepts. Narrative responses from both groups were overwhelmingly positive. After administration, the forms were returned to the committee for data compilation; 327 usable cases were available from the survey. Data were scrubbed, entered in SPSS and prepared for analysis.

One of the most commonly used techniques for the measurement of latent variables—factor analysis (FA)—was used for data analysis. With FA, large sets of (observed) variables can be reduced into smaller sets with each set (factor) representing an underlying construct or latent variable (DeCoster, 1998; Field, 2009; Hair, Black, Babin, & Anderson, 2010; Hinkin, 1998). Factor analysis therefore enables one to explore patterns or interrelationships in the data and perform data reduction if necessary (Afifi, May, & Clark, 2012; Bertsch & Pham, 2012). There are two types of factor analysis: exploratory and confirmatory.

Exploratory factor analysis (EFA) investigates the structure of the data without regard to the theory (Hair et al., 2006). It seeks to discover the nature of the construct(s) influencing a set of responses (DeCoster, 1998). On the other hand, Confirmatory factor analysis (CFA) is meant to test (and confirm) if, based on the theory developed beforehand, there exists a relationship between the various constructs. The next section presents the results, starting with EFA followed by CFA.

Results and Discussion

Exploratory Factor Analysis

The literature (see e.g., Field, 2009; Hair et al., 2003; Hair et al., 2006), suggests several steps to follow in conducting EFA on a dataset: partial correlations, Bartlett's test for sphericity, measuring sampling adequacy, factor extraction, factor loadings, communality, and factor rotation. These steps are given below and were followed for each of the proposed constructs (Student, Course, and Instructor).

Partial correlations. The partial correlations (off-diagonal values in the anti-image correlation matrix in SPSS) indicate the amount of unexplained correlation within a set of variables. The preferred values are within the ± 0.5 interval (Field, 2009; Hair et al., 2006); values outside the ± 0.7 range are considered unsuitable for factor analysis. As shown in Table 1, for the three constructs, none of the off-diagonal partial correlations exceed the ± 0.5 interval and therefore pass the partial correlation test.

Bartlett's test of sphericity. This determines whether the correlation between each of the survey items is statistically significant. The χ^2 values need to be significantly different from zero to lead to a conclusion that the items are measuring a single latent variable (Field, 2009; Hair et al., 2006). As shown in Table 2, all the χ^2 values are significant (at $p < 0.001$) indicating that the

Table 1

Anti-Image Correlation Matrix

Construct / Variable	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11
Student											
S1	0.881										
S2	-0.212	0.869									
S3	-0.079	-0.205	0.863								
S4	-0.095	-0.170	-0.178	0.850							
S5	-0.128	-0.121	-0.237	-0.186	0.870						
S6	0.001	-0.031	-0.198	-0.311	-0.198	0.846					
Course											
C7	0.926										
C8	-0.279	0.878									
C9	0.009	-0.364	0.884								
C10	-0.039	0.102	-0.337	0.888							
C11	-0.167	-0.212	-0.112	-0.248	0.906						
C12	-0.026	-0.024	-0.055	-0.046	-0.260	0.927					
C13	-0.170	-0.132	-0.007	-0.243	-0.213	-0.253	0.913				
Instructor											
I14	0.934										
I15	-0.279	0.950									
I16	-0.105	-0.074	0.928								
I17	-0.033	-0.122	-0.358	0.936							
I18	-0.020	0.029	-0.230	-0.199	0.949						
I19	-0.002	-0.079	0.073	-0.083	-0.037	0.959					
I20	-0.001	0.028	-0.090	-0.176	-0.099	-0.215	0.960				
I21	-0.112	-0.007	-0.007	-0.117	-0.105	-0.080	-0.128	0.954			
I22	-0.005	-0.102	0.135	0.081	0.031	-0.088	-0.045	-0.307	0.881		
I23	0.052	-0.016	-0.110	-0.100	0.079	0.019	-0.009	-0.113	-0.472	0.911	
I24	-0.337	-0.006	-0.103	-0.102	-0.197	-0.065	-0.144	-0.027	-0.149	-0.021	0.940

Note. "Var" indicates (corresponding) variable, e.g., Var1 under "Student" is S1; Var2 is S2, etc.

items in each respective construct pass the test of sphericity.

Measures of central adequacy. A commonly applied measure of sampling adequacy is the Kaiser-Meyer-Olkin (KMO) test. Values above 0.5 are said to be acceptable (Bertsch & Pham, 2012). In SPSS the general KMO for all variables included in the analysis is provided as well as the values for individual variables, that is, the diagonals of the anti-image correlation matrix (Field, 2009; Hair et al., 2006). The latter are subject to the same (0.5) threshold criteria as the

Table 2

KMO and Bartlett's Test

Student		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.862
Bartlett's Test of Sphericity	Approx. χ2	562.236
	Df	15.000
	Sig.	0.000
Course		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.902
Bartlett's Test of Sphericity	Approx. χ2	1261.133
	Df	21.000
	Sig.	0.000
Instructor		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.936
Bartlett's Test of Sphericity	Approx. χ2	2201.891
	Df	55.000
	Sig.	0.000

KMO values. As shown in Table 2, the KMO values for the three constructs exceed 0.8. Also, as depicted in Table 1, all the diagonal values exceed the 0.8. Thus, all the items in each of the three constructs both individually and collectively (within each of the three constructs) satisfied all the discussed measures indicating that all the variables (i.e., survey items) are worthy of factor analysis. Therefore, all the variables were subjected to the remaining steps in EFA.

Factor extraction. Principle components analysis (PCA) was chosen for factor extraction because of its common usage in the literature (Field, 2009; Hair et al., 2006). Additionally, it is psychometrically sound, and conceptually less complex than other methods (Bertsch & Pham, 2012). Kaiser's suggestion of "eigenvalues > 1"' was adopted in determining the appropriate number of factors for extraction (Field, 2009; Hair et al., 2003; Hair et al., 2006). As Table 3 illustrates, for each test of the three constructs, only one factor had an eigenvalue exceeding 1.

Factor loadings. The threshold for factor loadings to be considered significant varies with sample size. Field (2009, p. 644; citing Stevens, 2002), suggests a minimum (factor) loading of 0.298 for samples of 300 cases and 0.210 for samples of 600 cases or more. This would call for a factor loading between 0.210 and 0.298 for the sample at hand (n = 327). Hair et al. (2006, p. 128) is more specific, offering that samples above 250 require factor loadings of at least 0.35 while samples above 350 require factor loadings of 0.30 to be considered significant. Given the size of our sample (327 cases), a threshold of 0.325 was used to determine the retention of each survey item. As shown in Table 4 all the factor loadings exceed this 0.325 minimum threshold.

Communality. This is the portion of the variance attributable to the common factors in a variable, i.e., the shared variance explained by the extracted factor. Although there is no real rule-of-thumb to guide researchers, communalities and factor loadings are considered together in

Table 3

Course Construct: Total Variance Explained

Construct/ Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative
Student						
S1	3.169	52.818	52.818	3.169	52.818	52.818
S2	0.813	13.556	66.374			
S3	0.595	9.910	76.284			
S4	0.520	8.662	84.945			
S5	0.468	7.797	92.742			
S6	0.435	7.258	100			
Course						
C7	4.478	63.969	63.969	4.478	63.969	63.969
C8	0.613	8.751	72.720			
C9	0.574	8.205	80.925			
C10	0.453	6.473	87.396			
C11	0.341	4.865	92.261			
C12	0.289	4.129	96.389			
C13	0.253	3.611	100			
Instructor						
I14	6.542	59.477	59.477	6.542	59.477	59.477
I15	0.926	8.420	67.897			
I16	0.712	6.473	74.370			
I17	0.658	5.980	80.350			
I18	0.424	3.855	84.205			
I19	0.376	3.420	87.626			
I20	0.369	3.352	90.978			
I21	0.325	2.952	93.930			
I22	0.237	2.158	96.088			
I23	0.221	2.013	98.101			
I24	0.209	1.899	100			

determining whether to retain a variable in a factor solution (Field, 2009; Hair et al., 2006). None of the variables had exceptionally low communality.

Factor extraction (eigenvalues), factor loadings, and communalities were considered to determine what variables to drop from a given construct. There were no exceptional issues revealed in any of these steps. A final step (factor rotation) was used to determine the final rotated factor structure for the collective underlying dataset.

Factor rotation. For the project at hand, the goal was to create three separate and distinct constructs (i.e., Student, Course, and Instructor) for improving teaching and learning. Therefore, an orthogonal rotation method was used (see guidance by Hair et al., 2006). Equamax (version of orthogonal rotation) was selected as it attempts to simplify both the column and the row structures of the rotated matrix (Field, 2009; Hair et al., 2006), allowing for easier interpretation. Table 5 displays the rotated factor structure for the three constructs.

Table 4

Component Matrix[a]

Construct	Component
	1
Student	
S4	0.785
S3	0.777
S5	0.767
S6	0.730
S2	0.703
S1	0.579
Course	
C11	0.876
C13	0.838
C8	0.803
C9	0.792
C10	0.789
C12	0.749
C7	0.743
Instructor	
I24	0.882
I17	0.846
I21	0.812
I20	0.791
I16	0.783
I23	0.777
I14	0.774
I18	0.762
I22	0.733
I15	0.659
I19	0.628

Note. Extraction Method: Principal Component Analysis. [a]. 1 component extracted for each construct.

Hair et al. (2006), propose several rules of thumb for interpreting the factors derived with EFA. These steps include looking for items that cross load on more than one factor and retaining the higher factor loading while suppressing the lower factor loading; eliminating items that cross load with no clear and strong loading and reviewing the communalities; and using different rotational methods to, hopefully, find one that better defines the underlying structure (Hair et al., 2006). The process was applied to the exploration of the underlying factor structure for the data at hand. The resulting factor structure is very clean and clearly illustrates the presence of three distinct and mutually exclusive factors; where 'factor' reflects the presence of three distinct constructs of Student (Column 1, S1 through S6), Course (Column 2, C8 through C13), and Instructor (Column 3, I14 through I24).

Table 5

Rotated Component Matrix[a]

Variable	Component		
	1	2	3
S1			0.442
S2			0.666
S3			0.783
S4			0.783
S5			0.718
S6			0.738
C7		0.563	
C8		0.718	
C9		0.781	
C10		0.756	
C11		0.751	
C12		0.531	
C13		0.706	
I14	0.65		
I15	0.545		
I16	0.652		
I17	0.686		
I18	0.649		
I19	0.529		
I20	0.671		
I21	0.767		
I22	0.703		
I23	0.737		
I24	0.776		

Note. Extraction method: Principal Component Analysis. Rotation method: Equamax with Kaiser normalization. [a]. Rotation converged in 6 iterations.

Confirmatory Factor Analysis

Confirmatory factor analysis is meant to ratify that constructs included in a study qualify as such—that they are in fact distinct, autonomous constructs. Construct validity refers to the "extent to which a set of measured variables [indicators] actually represent the theoretical latent construct ..." (Hair et al., 2006, p. 707). The components of construct validity typically include convergent validity, discriminant validity, and face validity (Bertsch & Pham, 2012).

Discriminant validity is defined as the extent to which the measure of a construct does not correlate with other constructs that are supposed to be different (Hair et al., 2003; Singleton & Straits, 2005). For the purposes of this project, one would expect correlation between items due to the very nature of defining and operationalizing the constructs and collective instrument included in this study. Psychometric issues such as response bias, campus climate, and the discussion required therein are beyond the scope of this article. Therefore, construct validity tests will be limited to face validity and convergent validity.

Face validity can be assumed for several of the items given that those items were borrowed from existing instruments applied at other institutions of higher learning. The committee assessed face validity of those items created within the context of this project. The committee also queried students and faculty as part of the pilot study described earlier. As stated above, comments sought from both faculty and students during the pilot test were overwhelmingly in support of using these measures.

Convergent validity is often defined as the extent to which independent measures of the same concept correlate with each other (Hair et al., 2003; Singleton & Straits, 2005). Unfortunately, no known independent measure of each of the constructs (i.e., Student, Course, and Faculty) exists across the target sample, to which the data collected for this study could be compared to assess convergent validity. Therefore, this study limited the assessment of convergent validity to the analysis of the indicators contained in this instrument for each construct. The steps followed herein to assess convergent validity include factor loadings, variance extracted, and reliability (Bertsch & Pham, 2012).

When assessing convergent validity, Hair et al. (2006) suggest that factor loadings should be 0.5 or higher but ideally 0.7 or higher. As illustrated in Table 4, each of the factor loadings for all three constructs (Step 5 of EFA) was above the 0.5 threshold. All items satisfied the convergent validity benchmark. The test for variance extracted will be to determine whether the total variance extracted exceeds 0.5. This threshold indicates that more than 50% of the variance is explained by the observed variables, implying that less than 50% of the variance is attributed to error. Table 3 in Step 4 of the EFA illustrates that the three constructs of Student, Course, and Instructor all had more than 50% of the total variance extracted in a single factor (52.82%, 63.97%, and 59.48%, respectively), satisfying the said standard for variance extracted. Following Hair et al. (2006, p. 777), Cronbach's alpha coefficients were used to assess reliability, with a threshold of 0.6. For the Student construct, the Cronbach's alpha coefficient was 0.814; it was 0.904 for the Course construct, and, 0.929 for the Instructor construct. Thus, all three scales can be deemed reliable. Regarding convergent validity therefore, all the items contained in each of the three constructs converge onto a single respective and mutually exclusive construct. Furthermore, there is clear factor structure; both data-driven as shown by EFA and theory-driven as established by CFA. The proposed instrument passes known tests for reliability and validity.

Conclusion

This study was designed to create a new instrument that measures the student perception of learning as represented by the three concepts of Student, Course, and Instructor. As described in the beginning of this article, we acknowledge the limitations of our focus on only three latent variables: (i) the student concept—the student's perception of their own preparedness, engagement, and commitment to their own learning; (ii) the course concept—the course content, course organization, breadth of material covered, and the overall learning experience; and (iii) the instructor concept—instructor's knowledge of the course, rapport with the students, enthusiasm, and classroom behavior. We also acknowledge that we treated each of these constructs as a unidimensional construct even though there is evidence that each is complex and multidimensional including such concepts as *teaching for understanding, support from other students, organized studying, intention to understand, intrinsic and extrinsic academic motivation*, among other such constructs (see, e.g., Bertsch et al., 2016 and Parpala et al., 2013). Nevertheless, it was the intent of this research effort to develop a succinct instrument to measure students' perceptions of learning to include a reflection of their own engagement, the course content and organization, and finally the instructor's engagement.

A series of steps were taken to specify the content domain of the student perception of learning constructs, develop items to measure that domain, and determine the extent to which the developed items measure that domain. Furthermore, the instrument was tested within a higher education institution setting.

A review of the literature and existing SETs was combined with qualitative analysis of questionnaires from various academic institutions. This analysis indicated that the most influential factors were student's perception of their own preparedness and engagement, student's perception of the value of course, and student's perception of instructor's management of the classroom and course material. These three factors were formally defined, a set of items was written to measure each factor, and a survey was conducted to determine the extent to which the items reflected their intended constructs.

Operationalizing and scaling were conducted at the committee level using borrowed instruments and creating new survey items to include in a pilot test. The intended factor structure emerged. The result was an instrument that included preliminary demographic questions and 24 items scaled on a 5-point Likert scale. A copy of the instrument's 24 core items is provided in the appendix. The factor structure was initially determined through exploratory methods. The scales displayed convergent validity. All the items contained in the three constructs (student, instructor, and course) converge onto respective and mutually exclusive constructs. There is clear factor structure: both data-driven as tested by exploratory factor analysis and theory-driven as tested by confirmatory factor analysis. The proposed instrument is deemed sound, reliable, and valid. Adoption of the new instrument to measure student perceptions of learning using three distinct and operationalized latent constructs: Student, Course, and Instructor, is recommended.

For further study, student measures should be compared to external measures such as GPA and graduation rates. This will provide a different perspective on whether the proposed instrument reflects effective teaching and learning.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 385–486. doi.org/10.1037/0022-0663.82.2.219

Ackerman, D., Gross, L. B., & Vigneron, F. (2009). Peer observation reports and student evaluations of teaching: Who are the experts? *The Alberta Journal of Educational Research, 55*(1), 18–39. Retrieved from www.ajer.ca

Afifi, A., May, S., & Clark, V. A. (2012). *Practical multivariate analysis* (5th ed.) Boca Raton, FL: Taylor and Francis.

Aitken E. N. (2016). Grading and reporting student learning. In S. Scott, D. Scott, & C. Webber (Eds.), *Assessment in education. The enabling power of assessment* (pp. 231–260). Cham: Springer.

Beran, T. N., & Rokosh, J. L. (2009). The consequential validity of student ratings: What do instructors really think? *The Alberta Journal of Educational Research 55*(4), 497–511. Retrieved from www.ajer.ca

Berk, A. R. (2014). Should student outcomes be used to evaluate teaching? *Journal of Faculty Development 28*(2). 87–96. Retrieved from http://www.ronberk.com/articles.shtml

Bertsch, A., Nguyen, H., Clarke, G., Morman, M., Ofsthun, H., Peacock, A., Ondracek, J., & Saeed, M. (2016). Assessing Motivation in an Educational Setting, *Journal of International Doctoral Research, 5*(1), 70–96. Retrieved from https://www.idrcentre.org/journal

Bertsch, A., & Pham, L. (2012). A guide to multivariate analysis in cross cultural research. *Journal of International Doctoral Research, 1*(1), 97–121. Retrieved from https://www.idrcentre.org/journal

Bok, D. (1992). Reclaiming the public trust. *Change: The Magazine of Higher Learning, 24*(4), 13–21. doi:10.1080/00091383.1992.9937114

Catano, V. M., & Harvey, S. (2011). Student perception of teaching effectiveness: Development and validation of the evaluation of teaching competencies scale (ETCS). *Assessment & Evaluation in Higher Education, 36*(6), 701–717. doi:10.1080/02602938.2010.484879

Cathcart, A., Greer, D., & Neale, L. (2014). Learner-focused evaluation cycles: Facilitating learning using feedforward, concurrent and feedback evaluation. *Assessment & Evaluation in Higher Education, 39*(7), 790–802. doi:10.1080/02602938.2013.870969

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

DeCoster, J. (1998). *Overview of Factor Analysis*. Retrieved from http://www.stat-help.com/notes.html

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student grades and average ratings of instructional quality: the need for adjustment. *Journal of Educational Research, 97*(1), 35–40. https://doi.org/10.1080/00220670309596626

Entwistle, N. J., McCune, V., & Hounsell, J. (2003). Investigating ways of enhancing university teaching-learning environments: Measuring students' approaches to studying and perceptions of teaching. In E. De Corte, L. Verschaffel, N. Entwistle, & J. van Merrienboer (Eds.), *Unravelling basic components and dimensions of powerful learning environments* (pp. 89–107). Oxford: Elsevier Science.

Entwistle, N. J., & Ramsden, P. (1983). *Understanding student learning*. London: Croom Helm.

Ewell, P. T. (1991). Assessment and public accountability: Back to the future. *Change, 23*(6), 12–17.

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education 4*(1), 69–111.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry, & J. C. Smart (Eds.). *Effective teaching in higher education: Research and practice* (pp. 368–395). New York, NY: Agathon Press.

Freire, P. (1992). *Pedagogy of the Oppressed* (M. B. Ramos, Trans.). New York, NY: Continuum International Publishing Group. (Original work published 1921).

Field, A. P. (2009). *Discovering statistics using SPSS*. London: SAGE.

Golding, C., & Adam, L. (2016). Evaluate to improve: Useful approaches to student evaluation. *Assessment & Evaluation in Higher Education, 41*(1), 1–14. doi:10.1080/02602938.2014.976810.

Hair J. F., Babin, B., Money, A. H., & Samouel, P. (2003). *Essentials of business research methods*. Hoboken, NJ: Wiley.

Hair J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson.

Hair J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson.

Hinkin, T. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods 1*(1), 104–121. http://dx.doi.org/10.1177/109442819800100106

Johnson, V. E. (2003). *Grade inflation: A crisis in higher education*. New York, NY: Springer.

Kalender, I. (2015). Reliability-related issues in the context of student evaluations of teaching in higher education. *International Journal of Higher Education, 4*(3), 44–56. doi:10.5430/ijhe.v4n3p44

Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumption and missing conceptions about 'good' teaching. *Studies in Higher Education, 24*(1), 27–42. doi:10.1080/03075079912331380128

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 109*, 9–25. doi:10.1002/ir.1

Loes, N. C., Salisbury, H. M., & Pascarella, T. E. (2015). Student perceptions of effective instruction and the development of critical thinking: A replication and extension. *Higher Education, 69*(5), 823–838. https://doi.org/10.1007/s10734-014-9807-0

MacNell, L., Driscoll, A., & Hunt, N.A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. doi: 10.1007/s10755-014-9313-4.

Malhotra, N. K. (2007). *Marketing research: An applied orientation* (5th ed.). Upper Saddle River, NJ: Pearson.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77–95. doi:10.1111/j.2044-8279. 1982.tb02505.x.

Marsh. H. W. (2007). Students' evaluations of university teaching: dimensionality, reliability, Validity, potential biases and usefulness. In R. P. Perry & J. C. Smart, (Eds.). *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, Netherlands: Springer.

Marsh, H. W., & Roche, A. L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197. doi.org/10.1037/0003-066X.52.11.1187

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Marton, F., & Säljö, R. (1984). Approaches to learning. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning* (pp. 39–58). Edinburgh: Scottish Academic Press.

Nygaard, C., & Belluigi, D. Z. (2011). A proposed methodology for contextualized evaluation in higher education. *Assessment & Evaluation in Higher Education, 36*(6), 657–671. doi:10.1080/02602931003650037

Parpala, A., Lindblom-Ylänne, S., Komulainen, E., & Entwistle, N. J. (2013). Assessing students' experiences of teaching–learning environments and approaches to learning: Validation of a questionnaire in different countries and varying contexts. *Learning Environments Research, 16*(2), 201–215. https://doi.org/10.1007/s10984-013-9128-8

Patton, M. Q., Asibey, E., Dean-Coffey, J., Kelley, R., Miranda, R., Parker, S., & Fariss Newman, G. (2014, January 28). What is evaluation [Blog post]. Retrieved from https://www.eval.org/p/cm/ld/fid=1

Pennsylvania State University. (2003). University Faculty Senate Minutes (March 25, 2003) Appendix N: Annual grade distribution report. University Park, PA: Pennsylvania State University.

Powell, A. G., Farrar, E., & Cohen, D. K. (1985). *The shopping mall high school*. Boston, MA: Houghton Mifflin.

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education, 39*(8), 987–1001. doi:10.1080/02602938.2014.880777

Rojstaczer, S. (2002). Grade inflation at American colleges and universities. Retrieved from http://www.gradeinflation.com/

Rojstaczer, S., & Healy, C. (2010). Grade inflation at American colleges and universities. Retrieved from http://www.gradeinflation.com/

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record, 114*(7), 1–23. Retrieved from http://eric.ed.gov/?id=EJ1001967

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*, 1–16. doi:10.3389/fpsyg.2016.00110

Simmons, T. L. (1997). Student evaluation of teachers: Professional practice or punitive policy? *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(1), 12–19. Retrieved from http://jalt.org/test/sim_1.htm

Singleton, R. A., & Straits, B. C. (2005). *Approaches to social research*. (4th ed.). New York, NY: Oxford University Press.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Education and Education Research, 83*(4), 598–642. doi:10.3102/0034654313496870

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Titus, J. J. (2008). Student ratings in a consumerist academy: leveraging pedagogical control and authority. *Sociological Perspectives, 51*(2), 397–422. doi:10.1525/sop.2008.51.2.397

University of California, Los Angeles Office of Instructional Development [UCLAOID]. (2011). *The teacher's guide*. (5th ed.). Los Angeles, CA: University of California. Retrieved from https://ucla.app.box.com/v/teachers-guide.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191–212. doi:10.1080/0260293980230207

Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research in Higher Education, 35*(2), 251–266. doi.org/10.1007/BF02496704

Wines, A. W., & Lau, J. T. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decisions and its implications for academic freedom. *William and Mary Journal of Women and Law, 13*(1). Retrieved from https://scholarship.law.wm.edu/wmjowl/

Zelby, L. W. (1974). Student-faculty evaluation. *Science, 183*(4131), 1267–1270 doi:10.1126/science.183.4131.1267

Zikmund, A. G., & Babin, B. J. (2007). *Exploring marketing research*. (9th ed). Mason, OH: Thomson South-western.

Daniel George Ngugi is Associate Professor and Coordinator of Economics at Minot State University in the USA. His research interests include learning in higher education, environmental economics, managerial and business economics, and the economics of technology adoption.

Lisa Borden-King holds a Ph.D. with an emphasis on the philosophy of education from Indiana University. She has taught Preschool & Kindergarten in the past and has also taught at the college level for 25 years. Dr. Borden-King's research interests are in the areas of student learning, assessment of student learning, and reading instruction at all levels.

Draza Markovic is Associate Professor of Physics at Minot State University his areas of research interest include Rossby-Alfven waves, black holes, neutron stars, microlensing events, as well as teaching and learning.

Andy Bertsch holds a Doctor of Business Administration (DBA) degree in International Management and Leadership from the Henley Management College in England and an Advanced Postgraduate Diploma in Management Consultancy (APDMC) also from Henley. Dr. Bertsch actively researches and consults in the fields of leadership, international management, culture, entrepreneurship, strategic planning, and economic development.

Appendix: The SET Instrument (Core Items)

Student	SA	A	NT	D	SD	NA
1. I asked the instructor for feedback when I needed it						
2. I participated in class when appropriate and necessary						
3. I came prepared for class						
4. I attended class and related meetings						
5. I actively attempted to learn the material						
6. I completed all my assignments as required						
Please provide written comments to support your responses to questions 1-6:						
Course	SA	A	NT	D	SD	NA
7. The course was well organized						
8. The course materials were helpful and added to the learning experience						
9. The readings and assignments contributed to my learning						
10. The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view						
11. The assignments and classroom activities supported the course goals and objectives						
12. The work required for the course was appropriate for the credit given						
13. Overall, the course has been a worthwhile addition to my learning experience						
Please provide written comments to support your responses to questions 7-13:						
Instructor	SA	A	NT	D	SD	NA
14. The instructor effectively communicated ideas and information						
15. The instructor was well organized and prepared for class						
16. The instructor encouraged me to connect real world situations to the course when appropriate						
17. The instructor found ways to keep me interested and engaged						
18. The instructor treated me and my ideas and opinions with respect						
19. The instructor was available during posted office hours and/or by appointment						
20. The instructor encouraged class participation when appropriate						
21. The instructor provided useful and timely feedback						
22. The instructor explained grading criteria clearly						
23. The instructor applied grading criteria consistently						
24. Overall, the instructor was effective in promoting my learning in this course						

Please provide written comments to support your responses to questions 14-24:

Note. This form has been condensed to fit; the column headings abbreviated here were spelled out in the actual form: SA—*strongly agree*; A—*agree*; NT—*neutral*; D—*disagree*; SD—*strongly disagree*; NA—*not applicable*. An introductory and confidentiality statement as well as demographic questions (college year, etc.) preceded the (actual) form.