IJATE

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

Higher Education End-of-Course Evaluations: Assessing the Psychometric Properties Utilizing Exploratory Factor Analysis and Rasch Modeling Approaches

Kelly D. Bradley[1,*], Eric M. Snyder[2], Angela K. Tombari[1]

[1]University of Kentucky, Educational Policy Studies & Evaluation, Lexington, KY 40506, USA.

[2]University of Oklahoma, Jeanine Rainbolt College of Education, Department of Educational Leadership and Policy Studies, Norman, OK 73019, USA.

Abstract

This paper offers a critical assessment of the psychometric properties of a standard higher education end-of-course evaluation. Using both exploratory factor analysis (EFA) and Rasch modeling, the authors investigate the (a) an overall assessment of dimensionality using EFA, (b) a secondary assessment of dimensionality using a principal components analysis (PCA) of the residuals when the items are fit to the Rasch model, and (c) an assessment of item-level properties using item-level statistics provided when the items are fit to the Rasch model. The results support the usage of the scale as a supplement to high-stakes decision making such as tenure. However, the lack of precise targeting of item difficulty to person ability combined with the low person separation index renders rank-ordering professors according to minuscule differences in overall subscale scores a highly questionable practice.

Article Info

Received:
09.10.2015

Accepted:
11.10.2015

Keywords:
Course Evaluations; Rasch; Exploratory Factor Analysis; Psychometrics; Tenure

1. Introduction

Teaching, research, and service are the triad of higher education. As such, data-driven decision-making and accountability are two phrases that are commonly mentioned when attempting to measure how well these activities are performed by faculty. In particular, measuring teaching by way of end-of-course evaluations has been a fixed practice in higher education for many years (Guthrie, 1954; Otani, Kim, & Cho, 2011). A high-stakes assessment such as this is typically used as a measure of teaching quality and is linked to important and sometimes career changing decisions, such as determining whether promotion, tenure, and pay raises should be granted or during performance reviews.

When generating reports of the end-of-course evaluations, means and standard deviations are traditionally reported at the item level. This practice presupposes that the response scale functions in an approximately interval fashion such that the steps between scale points are equivalent. If these steps are found to not be equivalent, or if the scale is found to have any central pivot points that represent a dramatic difference in respondent assessment of teacher ability, then a reporting of means and standard deviations is not only faulty but also potentially very misleading. While this paper does not tackle all the issues

[*] Corresponding Author Phone: + 859-257-4923, Email: kdbrad2@uky.edu

2148-7456 /© 2016

associated with end-of-course evaluations, it does provide a critical assessment of the psychometric properties of a standard higher education end-of-course evaluation. Illustrating what the most common tools of practice measure well and not so well.

2. Theoretical framework

2.1. End-Of-Course Evaluation Debate

Student end-of-course evaluations in higher education are long withstanding, beginning with the first evaluations of the 1920s (Guthrie, 1954; Otani, Kim, & Cho, 2011). Originally intended as an impartial and scientific means to measure teaching performance, evaluations have become the topic of controversy between students and faculty, faculty and students, and faculty and administration (Calkins & Micari, 2010). These controversies have sparked contentious debates within the higher education community regarding the use of evaluations to measure and ultimately assess performance in the classroom and subsequently their use in making such high stakes decisions as deciding tenure and promotion. To begin the conversation, a critical psychometric review of the instrument is required.

In 1974, the American Association of University Profession (AAUP) released the Statement on Teaching Evaluation, which confirmed the importance of student input regarding quality teaching (Calkins & Micari, 2010). In response to the AAUP's assertion, many higher education institutions began to include student evaluations in personnel decisions (Thorne, 1980). Because institutions can utilize these evaluations to their advantage to monitor curricula and hold faculty accountable for student success, research into the concerns of end-of-course evaluations from a variety of perspectives has been conducted.

Prevalent findings are outlined here. Faculty were more skeptical of evaluations in comparison to administrators who believed that the responses to end-of-course evaluations represent an accurate description of effectiveness (Morgan, Sneed, & Swinney, 2003). There is evidence to suggest that end-of-course evaluations are a source of anxiety (Hodges & Stanton, 2007) and hostility (Franklin & Theall, 1989) for faculty. Faculty members often question the validity of student evaluations and the misuse of data (Beran, Violato, & Kline, 2007; Ory, 2001). A more recent article by Spooren, Brockx, and Mortelmans (2013) found many higher education stakeholders, including faculty, continue to question the usefulness and validity of student evaluations of teaching. Although the debate of score validity and reliability continues, in general, faculty members regardless of their institutional affiliation have grown accustomed to the practice of end-of-course evaluations of their teaching (Ewing & Crockford, 2008). Based on this general acceptance of employing the evaluation, a review of the usefulness and validity of the measure is again supported.

From the administrative perspective, quantitative rating of faculty teaching, used to document effective teaching, frequently take place through end-of-course evaluations. These quantitative summaries are typically reported with means and standard deviations (Laube, Massoni, Sprague, & Ferber, 2007). In general, administrators have a positive attitude toward course evaluation data and find it a useful source of information although validity concerns exist (Campbell & Bozeman, 2008). These concerns offer additional support for a psychometric study.

Research on student perceptions is limited. Students believe they are effective evaluators of teaching; however, they do not realize how the data collected can affect the faculty they evaluate, like through administrative decisions (Campbell & Bozeman, 2008; Wachtel, 1998). There is also evidence to suggest that students use the evaluation summaries to determine in which courses to enroll and which courses to avoid (Anderson, et. al, 2012).

Regardless of faculty, administrative, or student perceptions, a plethora of research suggesting that educators in higher education should make better use of end-of-course evaluation, and the collected data exists (Calkins & Micari, 2010; Campbell & Bozeman, 2008; Griffin & Cook; 2009; Otani, Kim, & Cho, 2011; Wolfer & Johnson, 2003). Cote and Allahar (2007) asserted that professional fear of student evaluations is a major contributing factor to grade inflation. Student evaluations of instructors were overemphasized in the tenure and promotion process (Wattiauz, et. al, 2010). As such, legitimate concerns about general bias and validity issues of student evaluations continue to exist. This potential for bias and validity issues, however, in no way renders the evaluations of teaching in higher education useless. If student end-of-course evaluations are to be a key component in the documentation of effective teaching, then institutions should be certain that the evaluations and subsequent collected data are functioning as expected and being analyzed appropriately. This study will differ from many studies by focusing on the reoccurring validity concerns by evaluating a newly revamped end-of-course evaluation instrument through both classical test theory and item-response theory approaches.

2.2. End-Of-Course Evaluation Items

The end-of-course evaluation utilized in the study was a reconstruction of the original used at the institution, with a careful review of the literature for items that should be added. Collecting course information is necessary. In consideration, there is evidence to suggest that certain courses at institutions of higher education require an elevated level of student-teacher and student-student interaction, whereas others are more individualistic and require little interaction (Brown & Green, 2003). Because variations in courses exist, it is critical that researchers, institutional administration, or whoever is in charge of the instrument, select items that accurately measure the individual responses regardless of course differences.

Research regarding what items to include within end-of-course evaluations is plentiful and varied. Student-evaluations of faculty are multidimensional, and the development of course evaluation instruments requires several items to be linked to specific measures that students consider important (Marsh & Dunkin, 1997). According to Marsh & Dunkin (1997) six categories commonly appear on end-of-course evaluations: (1) course content, (2) the instructor's communication skills, (3) student-teacher interaction, (4) course difficulty and workload, (5) assessment practices, and (6) student self-assessment (Cashin, 1995). In 2006, Bangert found the following four categories were critical to include in course evaluations (1) student/faculty interaction, (2) active learning, (3) time on task, and (4) cooperation among students. Other research has suggested that student learning, student sense of community, student engagement in learning, use of multiple learning techniques and prompt instructor feedback were critical categories to include in end-of-course evaluations (Kim Liu & Bonk, 2005). Furthermore, Kelly et al., (2007) suggested the following categories for inclusion within end-of-course evaluations; (1) instructor attributes, (2) course content, and organization as well as (3) grading and assessment. Lastly, a study by Hathorn & Hathorn (2010) found basic course information; measureable learning objectives, effective communication, and course organization were considered to be appropriate dimensions for end-of-course evaluations.

Although no single dimension is considered sufficient to validate student evaluations of faculty, researchers are aware that institutions need instruments that will allow them to gather information for a variety of courses quickly and economically (Spooren, Brockx, and Mortelmans, 2013). Since 2000, three peer-reviewed studies have been published that included instruments created for specific institutions (Barth, 2008; Cohen, 2005; Gursoy & Ubreit, 2005). The number of dimensions included in these studies were five (quality of instruction, course rigor, level or interest, grades, and instructor helpfulness) in Barth's (2008), two in Cohen's (course, teacher) (2005), and four in Gursoy & Umbreit (Organization, Workload, Instruction, and Learning) (2005) study. Specifically, the Cohen (2005) study identified two dimensions as critical to include in a students' overall evaluations to strengthen the validity of the measure. Variations of each of these dimensions, with additional sub-dimensions, were found to be essential within instruments developed specifically as larger scale (national) course evaluation measurements (Spooren, 2010; Mortelmans & Spoorten, 2009; Bangert, 2006; Toland & De Ayala, 2005; Kim, Liu & Bonk, 2005; Kelly et al., 2007; Hathorn & Hathorn, 2010; Marsh & Dunkin, 1997.) A review of each of the studies found items related to the course curriculum and material and the instructor. Therefore, the instrument developed and employed in this study was derived from the literature and focused on these two factors, largely due to the institution's existing tool, which had the only two items that were overall measures, one for instructor and one for course.

2.3. Psychometrics

Measurement is a component of the research process, especially in instrumentation, that is often taken for granted. When developing an instrument to measure certain traits, a researcher must be concerned with the quality of the instrument items and how the individual responds to those items. In order to eliminate these concerns, the reliability and validity of the instrument is commonly measured using psychometric techniques grounded in measurement theory. Psychometric theory offers two approaches to analyzing instrument data: classical test theory (CTT) and item-response theory (IRT).

While a Classical Test Theory is a common, widely used approach; here, a psychometric approach is applied. The item-response theory (IRT) approach to analyzing instrument items and respondents is based on item analysis and takes into consideration the chance of an individual's answering items right or wrong (Magno, 2009). It allows researchers to obtain an item characteristic curve for each item in the measure (Kaplan & Saccuzzo, 1997). The item characteristic curve describes the probability of responding correctly or incorrectly to an item given the ability of the individual. The Rasch (1960) model mathematically is equivalent to a one-parameter IRT model.

2.4. Purpose

The goal of this research is to assess a newly revamped end-of-course evaluation utilized at a Southeastern Research I higher education institution with particular attention given to item-level properties and overall validity. This investigation utilized a multi-tier analytical approach, with the steps being: (a) an overall assessment of dimensionality using exploratory factor analysis (EFA), (b) a secondary assessment of dimensionality using a principal components analysis (PCA) of the residuals when the items are fit to the Rasch model, and (c) an assessment of item-level properties using item-level statistics provided when the items are fit to the Rasch model. While not necessarily generalizable, results offer utility to the field of higher education. Results can support the interpretation of similar data and utility, or lack thereof of similar instruments at all higher education institutions.

3. Methods

3.1. Instrumentation

The instrument, with items mapped out in Table 3 and 5 below, was developed taking the institution's existing instrument, editing and vetting it through faculty council and the general faculty for feedback and edits. The literature was then reviewed and items that consistently appeared in the literature were added. Prior to use, faculty council, chairs, and the associate dean of research reviewed the instrument. Items are reviewed in more detail in the analysis.

3.2. Response Frame

Participants included students enrolled in eligible courses taught within college of education during the two summer sessions in 2013. Eligible courses met the following structure criteria: the course consisted of a practicum component of a lecture course with a different instructor from the lecture component or the course is labeled as a lab, a lecture, a seminar, or a distance-learning course. Courses that were irregular in structure, such as independent studies, were omitted from consideration, as these courses are not evaluated in a traditional fashion. Students were eligible for inclusion if they were enrolled in an eligible course past the final add/drop date for the summer session. Using the college Systems, Applications and Products (SAP) database, researchers identified 457 instances of enrollment past the final add/drop date in eligible courses. This resulted in 457 requests sent to 357 different people with the majority of individuals enrolled in only one course for the first summer session. Of these requests, 96 responses were returned, constituting a 21% response rate. Response return was slightly disparate as a higher response rate was found in courses traditionally identified as graduate courses (N=57, 25.79%) when compared to courses traditionally identified as undergraduate courses (N=39, 16.53%). The same protocol was used to identify instances of enrollment in eligible courses for the second summer session, leading to the identification of 676 requests, which were sent to 432 unique individuals. The return rate was slightly higher with 199 (29.44%) responses returned from 153 unique individuals. Responses rate for courses identified as graduate (N=158, 33.12%) exceeded response rate for courses identified as undergraduate (N=41, 20.60%), following the same pattern identified in the first summer session. Overall, 295 responses (26.04%) were collected across the two summer sessions for analysis.

3.3. Exploratory Factor Analysis

EFA was used as the foundation of the analysis due to a broad understanding of this procedure across most applied fields. This method offers an easily digestible assessment of dimensionality; however, EFA does have some shortcomings when it comes to assessing dimensionality. Most notably, EFA was primarily developed to look at cognitive abilities, which have the ability to be measured in such a way that the variables are continuous and normally distributed. Despite widespread application of EFA to Likert-type data, there are known weaknesses in applying an approach designed for continuous data to coarsely chopped ordinal data: (a) correlations assessed using Pearson's product-moment are attenuated due to floor and ceiling effects, (b) the number of factors to be extracted may be misleading, and (c) parameter estimates may be biased (Flora, LaBrish, & Chalmers, 2012). These issues are exacerbated if the ordinal data are also highly skewed, a common situation found in teacher end-of-course evaluation data. Although polychoric correlations are often used instead of product-moment correlations to analyze ordinal data, the polychoric correlation coefficient still assumes a *latent* normal continuous distribution. An additional weakness to this approach is that using polychoric correlation coefficients is still a limited-information technique in which only univariate and bivariate information is used to estimate the factor solution (Flora, LaBrish, & Chalmers, 2012).

3.4. Applying the Rasch Model

With concerns related to possible non-normality of items, a desire to assess the size of the steps between scale points, and recognition of the ordinal nature of the item responses, it was deemed preferable to use a polytomous Rasch model for the remainder of the analyses concerning this nascent scale. The Rasch model is a special case of the one-parameter logistic model from item-response theory (IRT) in which the discrimination parameter is set to a constant value of 1 rather than estimated from the data at hand. This demonstrates the pivotal difference between the Rasch family of models and the rest of the IRT models. In the Rasch models, the model is sacrosanct and represents characteristics desired in any measurement instrument: invariance of items/people, unidimensional measures, and independence of item and person parameters. The level to which these desired model characteristics can be applied to the data depends upon the fit of the data to the model. When operating from the mindset of IRT, the data are considered sacrosanct, with the model to be used (1-, 2-, or 3-PL) dependent on how well the model fits the data (Jaeger, 1977). Simply stated, in IRT it is a question of does the model fit the data? In a Rasch analysis, it becomes does the data fit the model?

For all Rasch analyses, the Andrich (1978) rating scale model was selected from the family of Rasch models for three reasons: (1) the same scale was used across items and should theoretically be consistent across items, (2) lower scale points, such as strongly disagree and disagree, were used less frequently, and (3) the Andrich rating scale has better stability with smaller sample sizes due to calculating across all items at the same time when compared with an estimation method that allows item thresholds to change across items, i.e. the partial credit model (Sick, 2009). The Andrich rating scale model formula is:

$$\log (P\mathit{nij} / Pni(j–1)) = B\mathit{n} – D\mathit{i} – Fj$$

where Pnij is the probability that person n encountering item *i* is observed in category *j*, B*n* is the "ability" measure of person *n*, D*i* is the "difficulty" measure of item *i*,. F*j* is the "calibration" measure of category *j* relative to category *j*-1, the point where categories *j*-1 and *j* are equally probable.

3.5. Data Analysis

Dimensionality. One of the key assumptions underlying a Rasch analysis is that of unidimensionality of measures. The teacher evaluation scale that has been developed includes three potential sub-dimensions: items pertaining to the curriculum and materials, items pertaining to the instructor, and items asking for reflection on the course overall. Items reflecting on the course overall were created with the intention of being potential explanatory variables for ratings as well as possible variables for use in assessing whether there is any differential item functioning present; therefore, only items pertaining to curriculum and instruction and items pertaining to the instructor were considered while assessing dimensionality. These two sets of items have the possibility of falling under one overarching dimension, such as course satisfaction, or being perceived as two distinct dimensions relating to unique aspects of the course experience. Dimensionality can be assessed with either classical test theory methods (e.g. reliability analysis and factor analysis) or Rasch methods (e.g., Rasch principal components analysis of residuals).

Using CTT methods, the common factor model was used to assess dimensionality with principal axis factoring as the preferred method of extraction due to robustness against non-normality, an expected issue when analyzing an end-of-course evaluation instrument. Preliminary analyses to assess the appropriateness of the data to analysis using dimension reduction techniques were examined using the Kaiser-Meyer-Olkin measure of sampling adequacy, in which values may range from 0 to 1 with values approaching 1 indicating stronger support for the existence of underlying factors, as well as Bartlett's test of sphericity, in which one seeks to reject the null hypothesis that the correlation matrix is an identity matrix. As the data were deemed acceptable for an EFA model, inspections of univariate indicators of normality, such as skewness and kurtosis values, as well as the application of Mardia's test of multivariate normality were used to assess the appropriateness of the factor extraction method. Retention of factors were determined by using multiple criteria: (1) a visual inspection of the Cattell (1966) scree plot, (2) examination of the eigenvalues using the Kaiser (1960) criteria, and (3) a review for theoretical coherence of the factor pattern and structure matrices. If more than one factor is suggested for retention using this method, the oblimin oblique rotation will be implemented, as the factors should be theoretically related: the oblimin method was selected purely for purposes of familiarity, as all oblique rotation methods produce similar results (Osborne, Costello, & Kellow, 2008).

A subsequent analysis to check dimensionality was performed using a Rasch PCA of item residuals. A PCA residual analysis began with the hypothesis that the residuals are error or random noise. The residuals were then grouped to explain as much of the remaining variance as possible in a contrast to see if the null hypothesis should be rejected. This is similar to the methods used in EFA to decide upon the number of components, such as the Cattell (1966) scree plot or the Kaiser (1966) criterion, as both EFA and a PCA of residuals attempt to assess whether subsequent eigenvalues exceed the value expected purely by chance or random noise (Raîche, 2005). Although similarities exist between EFA and the Rasch PCA of residuals, the output is interpreted differently. Common factor analysis seeks to optimize the factor structure solution by considering many indicators, such as (a) variance explained, (b) reduction of cross loadings, and (c) commonalities. The Rasch PCA analysis of residuals functions from one simple hypothesis: The residuals are just random noise. To attempt to nullify this hypothesis, the residuals are grouped in a way to explain a maximum variance in the first contrast. Although no absolute cutoff values determine when the items residual contrast denotes a true secondary dimension, rules of thumb help guide this decision: (a) the first contrast should explain at least the amount of variance in two items to be considered as a secondary dimension, (b) the value of the first contrast should exceed the value of the first contrast in randomly generated simulated data that conforms to the same data structure (i.e., number of respondents, number of items, and number of response options), and (c) the contrast should make theoretical sense (Linacre, 2014a). When the second contrast is large and theoretically sound, item statistics will be best parameterized by assessing each sub-dimension separately.

Item misfit. The Rasch model offers indices to assess individual item fit to the model in the form of infit and outfit mean squares. These two indices differentially weight people's responses: Infit offers an information-weighted measure of misfit, meaning that person responses with an estimated overall ability near the estimated item difficulty are given a greater weight, whereas outfit is an outlier-sensitive measure of misfit that reflects surprising responses by individuals with ability estimates disparate from the item's estimated difficulty. For both infit and outfit mean-squares, values of 1 are expected, and the values may range from 0 to infinity (Wright & Linacre, 1994).

As the Rasch model is a stochastic model, some randomness is expected when persons encounter the items. Thus, there are two ways in which an item (or person) can demonstrate misfit: underfit and overfit. Underfit occurs when too much randomness is perceived in an item, indicated by mean-square values greater than 1; in contrast, overfit occurs when response patterns to an item are overly predictable, indicated by mean-square values less than 1. Both of these occurrences are worth further attention, but overfit is a more pressing issue than underfit. Overfit can actually degrade the measurement capabilities of a scale, but underfit indicates an item that may not contribute much to the measurement abilities of the scale but is unlikely to actually reduce the measurement properties of the scale. One concern associated with underfit is the possible artificial inflation of person and item reliability indices. In general, items with mean-square values between .6 and 1.4 are deemed acceptable for a survey; however, it is important to note that there are no true absolute values that deem an item unacceptable (Wright & Linacre, 1994).

Variable map. A variable map offers the unique opportunity to place item difficulty and person ability on the same metric to assess appropriate matching of item difficulty and person ability. In an analogy derived from the physical sciences, a mismatch of items to persons can be similar to trying to measure the temperature of a roast with a thermometer intended to measure human temperatures. If the thermometer is not calibrated for the range desired, the precision of the measurement is diminished. Another way to think of this mismatch is to imagine and instructor giving calculus students a multiplication test. Due to the relative easiness of this assessment for individuals so advanced in mathematics, the instructor can tell that test-takers have mastered multiplication, but the instructor is unable to rank the individuals in terms of ability. If the instructor then gave these same students a calculus assessment on a current topic, the individual would be able to rank students on their ability by looking at the estimated difficulties of correct and incorrect items for each student.

Reliability estimates (person and item). Rasch reliability indices exist on the same metric as traditional CTT reliability indices, such as Cronbach’s alpha, and range from 0 to 1 with larger values demonstrating superior reliability. The reliability index reflects the reproducibility of the person or item order from most agreeable to least agreeable persons or most to least easily endorsed items. Two preconditions are necessary for acquiring a high reliability index: (1) the group (people/items) for which reliability is being estimated has to have a wide distribution across ability/difficulty, and (2) the opposing group (items if measuring people reliability and vice versa) needs to have sufficient length. Attaining a high person reliability index requires a relatively wide person ability distribution and sufficient items, and item reliability requires a relatively wide item difficulty distribution and sufficient people to attain a high reliability index. The appropriate matching of item difficulty to person ability is also necessary for achieving high reliability indices (Linacre, 2014b).

Step difficulties. The Andrich rating scale treats each transition across categories as an independent dichotomy. To guarantee that these transitions can actually be perceived as independent dichotomies, step difficulties must meet certain parameters: namely, step difficulties should advance by at least a bit more than one logit for a four category rating scale and no more than five logits for optimal measurement precision (Linacre, 2002). Applying all of the techniques and methods above, the evaluation results psychometrically sound and even more important from a practical standpoint, trustworthy to those individuals being evaluated.

4. Results

4.1. Exploratory Factor Analysis

Initial assessments for the appropriateness of attempting an explanation of the items with common factors supported the utility of performing an EFA. The Kaiser-Meyer-Olkin measure of sampling adequacy was found to be .92, suggesting that a great deal of the variance can be explained by underlying factors (e.g., small partial correlations after the shared variance is parceled out); furthermore, Bartlett’s test of sphericity, $\chi^2(78) = 3679.343$,

$p < .01$, suggested that the items are acceptable for factor analysis with enough covariation to allow reduction to a smaller set of factors.

The choice of extraction method for exploratory factor analysis (EFA) depends on whether the data meets the underlying assumptions of the extraction method. Preliminary review of univariate statistics was performed in light of suggested guidelines by West, Finch, and Curran (1995), in which absolute values of skew less than 2 and absolute values of kurtosis less than 7 indicated acceptable levels of univariate normality (see Table 1). Using this guideline, all items except Instructor3 "Instructor returned exams and/or papers in a timely manner" violated acceptable levels of skewness, but only one item, C&M6 "Exams were connected to course content", violated acceptable levels of kurtosis. This is an expected result because the data are technically ordinal rather than interval and used a scale with few scale points (4-point Likert-type ranging from *strongly disagree* to *strongly agree* with no midpoint).

Table 1. Univariate Normality Assessment for Items

	Skewness		Kurtosis	
Variable	Skew	*SE*	Kurtosis	*SE*
C&M1	**-2.270**	0.150	5.454	0.300
C&M2	**-2.162**	0.151	4.885	0.302
C&M3	**-2.299**	0.151	5.239	0.302
C&M4	**-2.068**	0.152	4.506	0.303
C&M5	**-2.381**	0.151	6.079	0.302
C&M6	**-2.493**	0.175	**7.522**	0.349
Instructor1	**-2.253**	0.151	5.141	0.302
Instructor2	**-2.393**	0.155	5.789	0.309
Instructor3	-1.995	0.156	4.162	0.310
Instructor4	**-2.329**	0.153	5.168	0.304
Instructor5	**-2.097**	0.152	4.361	0.302
Instructor6	**-2.038**	0.153	3.762	0.304
Instructor7	**-2.502**	0.155	6.178	0.308

Note. Bolded skew or kurtosis values indicate a violation of acceptable levels of skew and/or kurtosis for that item.

Univariate normality is a necessary, though not sufficient, precondition to multivariate normality; therefore, with the rejection of univariate normality, multivariate normality is not expected (DeCarlo, 1997). However, since univariate normality was assessed using rules of thumb rather than statistical analysis, Mardia's test for multivariate normality was also used for purposes of statistical conclusion validity. The results of this statistical test indicated a Mardia value of 667.5776, which suggests that the kurtosis of the empirical distribution differs significantly, $p < .001$, from the expected distribution of multivariate normal data. Thus, multivariate normality is an untenable assumption.

Preliminary analysis of the data produced a first factor that explained 78.54% of the variance (eigenvalue of 10.33) and a second factor that explained 9.35% of variance (eigenvalue of 1.33) after oblique rotation. All subsequent eigenvalues were less than .28 in value. Usage of the Kaiser (1966) criteria (or eigenvalue-greater-than-one rule) and examination of the Cattell (1960) scree plot (Figure 1) both lead to the same conclusion, a two-factor solution, explaining 88.89% of the variance in these items, appears to fit the data well.

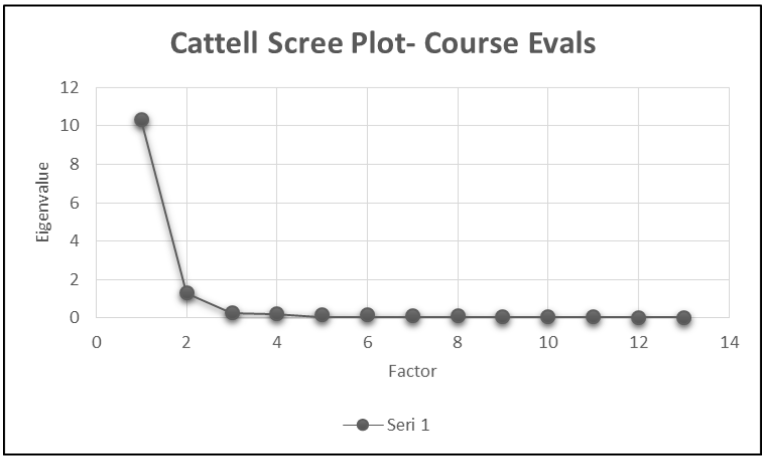


Figure 1. Cattell scree plot for Eigenvalues of C&M and Instructor items.

The two-factor solution was then examined for theoretical coherence. As can be seen in Table 2, the factor loading matrix, the items split based upon whether the items referred to the curriculum and materials subdomain or to the instructor subdomain. This fit with the development of the items and the content areas that were intended to be targeted; furthermore, cross loadings were very small across subdomains (Keep in mind that the two factors were allowed to correlate). The first factor consisted of the seven instructor items, and loadings ranged from .810 for instructor item 1, "The instructor was accessible when I needed help with course material," to .995 for instructor item 4, "The instructor advanced my knowledge of the subject." The largest cross loading from the curriculum and materials (C&M) items showed a loading of .075 for C&M item 3. The second factor consisted of the six C&M items, and factor loadings ranged from .864 for C&M item 3, "Course content was organized", to .987 for C&M item 6, "Exams were connected to course content." The largest cross loading from the instructor items showed a loading of .131 for instructor item 1. Factors 1 and 2 were found to have a .777 inter-factor correlation, showing them as strongly related subdomains.

Table 2. Factor Loadings after Direct Oblimin Rotation

	Factor	
Item	1	2
C&M1	-.069	**.985**
C&M2	.017	**.941**
C&M3	.075	**.864**
C&M4	.057	**.892**
C&M5	.015	**.929**
C&M6	-.023	**.987**
Instructor1	**.810**	.131
Instructor2	**.975**	-.026
Instructor3	**.812**	.102
Instructor4	**.995**	-.085
Instructor5	**.958**	-.028
Instructor6	**.942**	.003
Instructor7	**.930**	.027

Note. Bolded pattern loadings indicate that the item is retained on this factor.

4.2. Rasch PCA of Item Residuals

Despite the EFA suggesting a two factor solution, the high factor intercorrelation, .777, left open the possibility that a one-dimensional model, including both curriculum and instructor items, may demonstrate acceptable fit for a Rasch analysis. The central question is whether these two subscales are *different enough* to warrant independent investigation as separate unidimensional constructs; therefore, a one-dimensional model including both subscales was initially fit using the Andrich rating scale in Winsteps.

In the case of this particular PCA of residuals, the eigenvalue of the first contrast had a value of 3.8, suggesting that this eigenvalue shows the strength of about four items. This exceeds the base value of 2 needed to posit a secondary dimension contained within the data structure (Linacre, 2014a). When this first contrast is examined, the items group according to the proposed subscales, suggesting that these two subscales are distinct enough to warrant separate Rasch analyses. To assess this decision, expected eigenvalues and explained variance that can be attributed to noise was calculated by simulating 10 data files to match the response set, number of items, and number of respondents. On average, the first contrast from the simulated data showed an eigenvalue of 1.4 with a percentage of explained variance equaling 3.36%. This can be compared to the empirical results from the actual data that showed a real eigenvalue of 3.8 and the percentage of explained variance as 13.3% for the first contrast. Based on this information, the two subscales were analyzed separately for all remaining analyses, as they are different enough to preclude a combined unidimensional analysis.

4.3. Rasch Analysis of the Curriculum and Materials Subscale

Item misfit. All but two of the items on the Curriculum and Materials subscale demonstrate adequate fit when considering mean-square infit and outfit statistics as well as standardized *z*-scores (see Table 3). The most concerning issues involve underfit, meaning that mean-square infit values are greater than approximately 1.4, as this can reduce the quality of measurement due to too little predictability in respondents' use of the item. Item C&M3 “Course content was organized” had a mean-square infit statistic of 1.58, which indicated that the item was being answered unexpectedly by individuals to whom this item was appropriately targeted. As this item did not function in accordance with the rest of the items on the scale, no value was added to the overall scale by including this item. Item C&M3 should be indexed for either rewording or deletion during future development of this scale.

Item C&M6 “Exams were connected to course content” displayed the opposite fit issue in that it was overfitting with a mean-square infit value of .48 and mean-square outfit value of .35, indicating that this item performed too predictably to provide useful additional information for this scale. Although this item could be considered for deletion or revision, overfit will not actually damage the scale measurement properties as a whole, and the item is theoretically important to include. If students indicated that exams were not connected to course content, then this would be extremely valuable information for the college to know; therefore, future iterations of this scale should retain this item despite displaying some misfit.

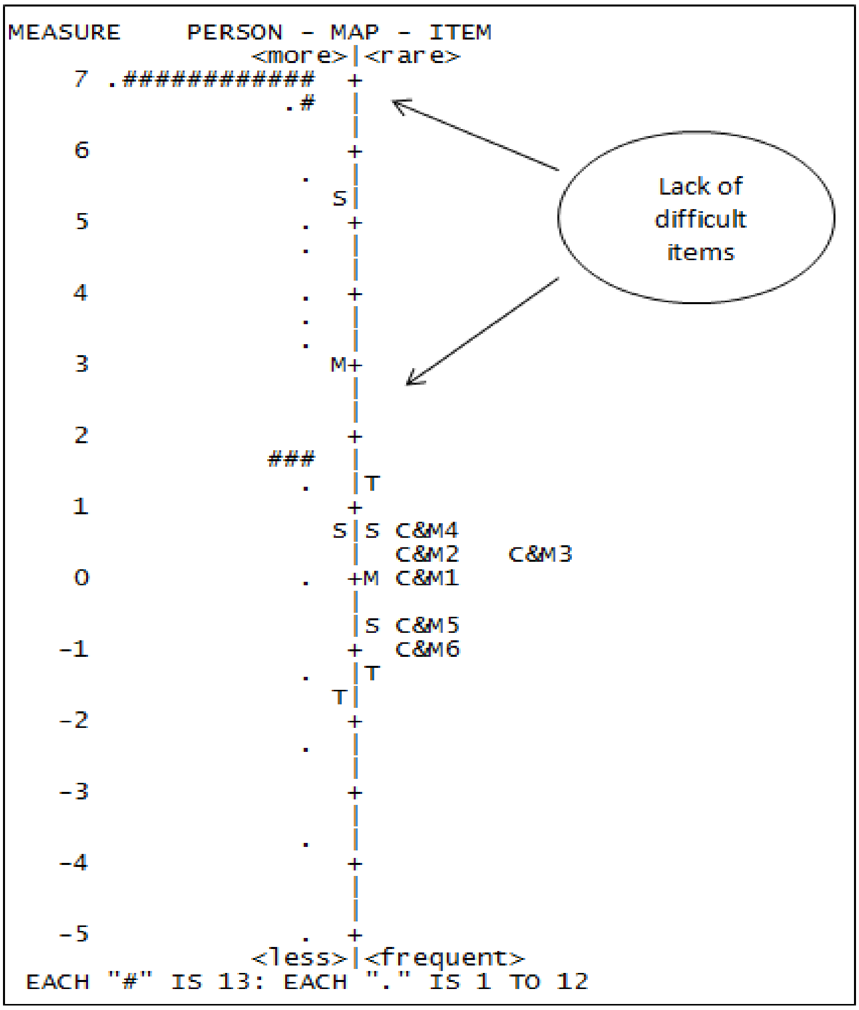
Table 3. Fit Statistics for Curriculum and Material Items

Item		Infit		Outfit	
Number	Stem	MNSQ	ZSTD	MNSQ	ZSTD
C&M3	Course content was organized.	**1.58**	**2.6**	1.35	1.2
C&M1	Objectives for this course are clearly stated.	1.26	1.3	1.10	.5
C&M2	Assignments contributed to my understanding of the material.	.82	-1.0	.74	-.9
C&M5	Assignments were connected to course content.	.80	-1.0	.61	-1.5
C&M4	The resources (texts, articles, videos, etc.) used in this course contributed to my learning.	.79	-1.2	.65	-1.4
C&M6	Exams were connected to course content.	**.48**	**-2.5**	**.35**	**-2.6**
M		.96	-.3	.80	-.8
SD		.36	1.7	.33	1.3

Note. Bolded values are outside the recommended values, suggesting misfit.

Variable map. The variable map shows a lack of congruence between item difficulty and person ability (see Figure 2). The C&M items are easily endorsed by the individuals included in this sample. Out of the six items, the easiest item to endorse is C&M6, "Exams were connected to course content", and the most difficult item to endorse is C&M4, "The resources (texts, articles, videos, etc.) used in this course contributed to my learning." These items spanned a range of approximately two logits, suggesting that items are somewhat variable in difficulty. With this in mind, the common practice of taking an average across items to determine an overall score is questionable. Another important note from this map is that these items cannot differentiate among instructors with much precision on the upper end of the scale: As long as an instructor is getting responses of "agree" or "strongly agree," the instructor is performing well. These items should be considered as more of a pass-fail measure than as a way to differentiate the quality of instruction by ranking instructors on average scores.

Reliability estimates (person and item). As is expected from the mismatch between item difficulty and person ability overall, person reliability for the C&M subscale is calculated as .65, which suggests that it is possible to discriminate between only one to two levels for persons. This could be improved by increasing the number of categories on the response scale, presuming that the respondents can discriminate consistently across more scale points, or adding items that are more difficult to endorse to better match the higher person ability on this scale. Item reliability (.77) is higher than the person reliability, but this still only suggests the ability to discriminate across two to three levels of items. In order to improve item reliability, items that are more difficult to endorse could be generated and included in future iterations of this scale.



```
MEASURE     PERSON - MAP - ITEM
                   <more>|<rare>
    7  .############  +
                   .#  |
                       |
    6                  +
                     . |
                      S|
    5                .  +
                     .  |
                        |
    4                .  +
                     .  |
                     .  |
    3                  M+
                        |
                        |
    2                   +
                  ###   |
                    .   |T
    1                   +
                       S|S C&M4
                        |   C&M2     C&M3
    0                .  +M C&M1
                        |
                        |S C&M5
   -1                   +  C&M6
                     .  |T
                       T|
   -2                   +
                     .  |
                        |
   -3                   +
                        |
                     .  |
   -4                   +
                        |
                        |
   -5                .  +
                   <less>|<frequent>
  EACH "#" IS 13: EACH "." IS 1 TO 12
```

Figure 2. Variable (person/item) map for Curriculum and Materials items.

Step difficulties. As can be seen in Table 4, the threshold between one (strongly disagree) and two (disagree) occurs at -3.14, and the threshold between two (disagree) and three (agree) occurs at -1.73. This yields a step difficulty of 1.41 logits, which is close to optimal for a scale with four response categories. The issue arises when looking at the threshold between three (agree) and four (strongly agree), which occurs at 4.88 logits, thereby rendering a step difficulty of 6.61 logits. Although this is adequate for separate dichotomies, measurement precision was lost because the step difficulty is too broad. This is similar to trying to measure people's heights with a ruler only marked in feet: Many distinctions will be lost between individuals who range in height between five and six feet.

Table 4. *Summary of Category Structure for C&M Items*

Category	Observed		Observed	Sample	MNSQ		Andrich	Category
Label	Count	%	Average	Expectations	Infit	Outfit	Threshold	Measure
1	51	3	-2.79	-3.11	1.44	1.54	NONE	(-4.39)
2	22	1	-.11	-.30	1.14	.85	-3.14	-2.44
3	343	23	2.20	2.25	.85	.78	-1.73	1.58
4	1071	72	5.67	5.60	.96	.78	4.88	(5.98)

4.4. Rasch Analysis of the Instructor Subscale

Item misfit. All items on the Instructor subscale demonstrated adequate fit when considering mean-square infit and outfit statistics as well as standardized z-scores (see Table 5). One item, Instructor item 3, displayed a standardized infit z-score (2.3) that exceeded the desired range of ±2; however, the infit mean-square value was within acceptable range (1.38) as it is below the 1.4 cutoff guideline used in evaluating the items. Instructor item 3 demonstrated more unpredictability than desired in responses by individuals whose ability level was targeted by the item difficulty. With only one of the four fit statistics flagging the item as misfitting, the researchers retained this item, "Instructor returned exams and/or papers in a timely manner," for the added theoretical diversity. Overall, this subscale showed adequate item fit.

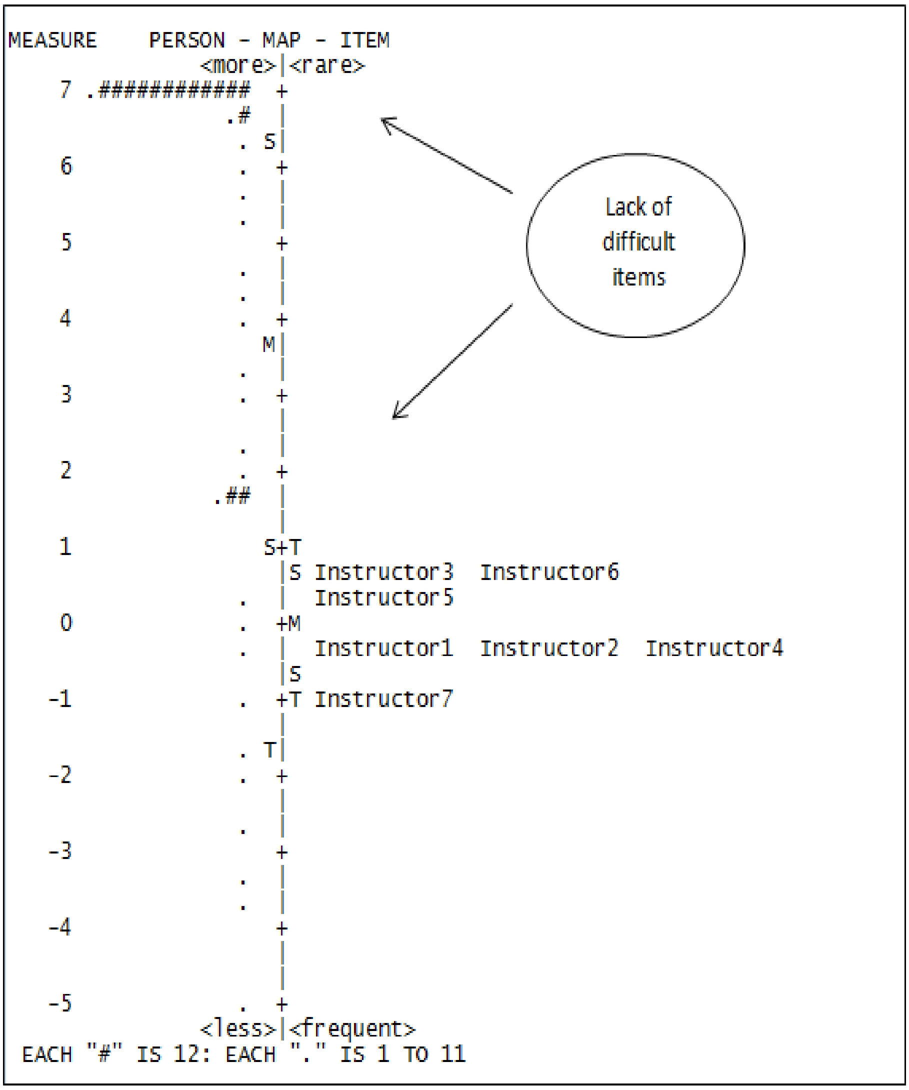
Table 5. Fit Statistics for Instructor Items

Item		Infit		Outfit	
Number	Stem	MNSQ	ZSTD	MNSQ	ZSTD
3	Instructor returned exams and/or papers in a timely manner.	1.38	***2.3***	1.23	1.1
1	Instructor was accessible when I needed help with course material.	1.03	.3	.94	-.2
6	Instructor presented course material in an effective manner.	1.03	.3	.85	-.7
4	Instructor advanced my knowledge of the subject.	1.01	.1	.81	-.8
7	Instructor made me feel comfortable with asking questions and expressing my ideas.	.87	-.7	.67	-1.3
5	Instructor incorporated current developments in the area of study.	.85	-1.0	.74	-1.3
2	Instructor was prepared for class.	.75	-1.4	.64	-1.6
M		.99	.0	.84	-.7
SD		.19	1.1	.19	.9

Note. Bolded values are outside the recommended values, suggesting misfit.

Variable map. The person-item map showed the same trends as found in the C&M subscale: Person ability exceeded item difficulty (see Figure 3). Essentially, respondents very easily endorse these items. The easiest item for the respondents to endorse was Instructor item 7, "Instructor made me feel comfortable with asking questions and expressing my ideas". The two most difficult items to endorse were Instructor item 3, "Instructor returned exams and/or papers in a timely fashion," and Instructor item 6, "Instructor presented course material in an effective manner." Instructor items 1, 2, and 4 appear to be redundant as they are estimated at the same level of item difficulty; however, these items should be retained for theoretical purposes because they address three different components of teacher instruction: accessibility of the instructor, preparedness for class, and advancing student knowledge in the topic area.

Overall, the mismatch between item difficulty and person ability, or willingness to endorse the item, suggests that these items can be used as a basic litmus test to assess whether the instructor is performing acceptably or intervention is necessary. Attempting to rank order instructors based on some scoring rule would be questionable because the mismatch leads to reduced precision of course instructor ability scores.



```
MEASURE    PERSON - MAP - ITEM
                  <more>|<rare>
    7 .############  +
                  .#  |
                   . S|
    6              .  +
                   .  |
                   .  |
    5                 +
                   .  |
                   .  |
    4              .  +
                     M|
                   .  |
    3              .  +
                      |
                   .  |
    2              .  +
                 .##  |
                      |
    1                S+T
                      |S Instructor3  Instructor6
                   .  |  Instructor5
    0              .  +M
                   .  |  Instructor1  Instructor2  Instructor4
                      |S
   -1              .  +T Instructor7
                      |
                   . T|
   -2              .  +
                      |
                   .  |
   -3                 +
                   .  |
                   .  |
   -4                 +
                      |
                      |
   -5              .  +
                  <less>|<frequent>
 EACH "#" IS 12: EACH "." IS 1 TO 11
```

Figure 3. Variable (person/item) map for Instructor items.

Reliability estimates (person and item). Both person reliability (.79) and item reliability (.76) were slightly lower than desired and discriminated between two to three levels of persons and items. In order to improve these values, items that are more difficult to endorse need to be constructed and added to this scale for future iterations.

Step difficulties. As can be seen in Table 6, the threshold between categories 1 and 2 occurred at -3.52 logits, and the threshold between categories 2 and 3 occurred at -1.39 logits. This yielded a step difficulty of 2.13 logits, which is adequate for presupposing two independent dichotomies. Similar to the C&M scale, the problem occurred at the transition between “agree” and “strongly agree”. The threshold for this transition is at 4.91 logits, which rendered a step difficulty of 6.30 logits. Such an increase will render a decrease in overall test information in the person ability level captured between item categories 3 and 4.

Table 6. Summary of Category Structure for Instructor Items.

Category	Observed		Observed	Sample	MNSQ		Andrich	Category
Label	Count	%	Average	Expectations	Infit	Outfit	Threshold	Measure
1	71	4	-3.10	-3.09	.91	.77	NONE	(-4.70)
2	38	2	- .45	-1.01	1.41	.92	-3.52	-2.46
3	380	22	2.58	2.70	.97	.89	-1.39	1.77
4	1275	72	6.08	5.99	.88	.75	4.91	(6.01)

5. Discussion

The literature currently suggests that higher education should make better use of end-of-course evaluations and subsequent data (Calkins & Micari, 2010; Campbell & Bozeman, 2008; Griffin & Cook, 2009; Otani, Kim, & Cho, 2011; Wolfer & Johnson, 2003). Much of this stems from faculty (Beran, Violato, & Kline, 2007) and administrator (Campbell & Bozeman, 2008) disappointment in the reliability and validity of current course evaluation measures. Given the faculty and administrator views of end-of-course evaluations, it is suggested that both groups receive training in how to administer the measure and interpret the results prior to its use in evaluation. Specifically, those analyzing and interpreting the data need to see beyond the tradition presentation of means and standard deviations. As illustrated above, a psychometric analysis offers greater insight and details about items and respondents.

While this study does not cover such issues as number of evaluations needed, students per implementation and the predictive validity of the measure with such outcomes as course outcomes, it does provide a strong foundation for the general critique of fit and function of items. The result of this research contributes to the greater body of evaluation literature and offers specific guidance in the realm of end-of-course evaluation literature, specific to the argument of validity. The ability to measure and evaluate faculty performance and the quality of higher education course curriculum and material is challenging. The use of EFA and Rasch modeling helped ameliorate this challenge because it allowed for the comparison of persons, items, and the underlying subscales that may influence the measure.

The items from the revamped end-of-course evaluation behaved in the expected manner. The EFA confirmed a two-factor solution categorizing items into Curriculum and Materials and Instructor. This is not surprising giving the literature and the construction of the instrument. However, the high factor intercorrelation, .777, left open the possibility that a one-dimensional model that includes both curriculum and instructor items may demonstrate acceptable fit for a Rasch analysis. Based on this information, a Rasch analysis was conducted that included both subscales and confirmed that they were different enough to complete a separate unidimensional analysis. The consistent identification of the two subscales (i.e., curriculum and materials and instructor) using both EFA and Rasch was appropriate given the question wording and support from the literature to include these components in the measure (Banger, 2006; Cashin, 1995; Hathorn & Hathorn, 2010; Kelly et al., 2007).

The separate Rasch analyses of the C&M scale revealed that C&M3 “Course content was organized” needs to be reworked or deleted before institutions use it for course evaluation purposes. The item map revealed the use of averages across items to determine an overall score could be problematic. The person reliability (.65) and item reliability (.77) were moderately low; therefore, it is recommended the instrument be modified to include more difficult items to endorse, more response categories, and an uneven response scale that allows greater differentiation on the “agree” side of the scale (i.e., strongly disagree, disagree, slightly agree, moderately agree, and strongly agree). A disadvantage to such a modification is that comparability to other end-of-course evaluation forms and prior end-of-course evaluations is lost by altering the scale.

All seven items on the Instructor subscale fit adequately. The person-item map showed that the items are easily endorsed. Instructor7 "Instructor made me feel comfortable with asking questions and expressing my ideas" was the easiest to endorse, but Instructor3 and Instructor6 were the most difficult to endorse. Both person reliability (.79) and item reliability (.76) were slightly lower than desired. To improve these values, items that are more difficult need to be constructed and added to this scale for future iterations. Again, this measure could be used as a pass/fail to identify when a conversation or intervention with a faculty member is necessary. Basically, it could be used as a screening instrument, where scores below a set mark are flagged for intervention by administration.

Without changing the measure, the usage of the scale as a supplement to high-stakes decisions, such as tenure, must be considered. The lack of precise targeting of item difficulty to person ability combined with the low person separation index renders rank-ordering professors according to minuscule differences in overall subscale scores a highly questionable practice. Instead, these items should be taken as an indicator that the course instruction is meeting a set of standards with any average score of 3.0 or above in a subscale deemed acceptable.

Future studies should investigate whether the instrument remains stable in various course settings by comparing findings (e.g. lecture evaluations versus distance education evaluations). If modifications to the current measure were performed, then an institution of higher education might use the resulting data to establish a benchmark to determine if there is a need for reinvention of course curriculum, workshops, or professional development for faculty. For example, an institution could use items 1-7 of the Instructor subscale as focus areas for brown-bag lunch events.

An institution of higher education could also use the measure to establish a required minimum score on end-of-course evaluations for all faculty members. For those faculty members who do not meet the minimum score, it would be possible to provide a teaching intervention. Again, the intervention could be tailored to the needs of the students given the data derived from administering the measure. For example, if assignments do not contribute to the students' understanding of the material (C&M2), then a workshop focusing on assignment creation can be provided.

It is imperative that higher education institutions use sound methods when determining the effectiveness of end-of-course evaluations. Unfortunately, it is extremely difficult for institutions to rectify poor or inappropriate course evaluation practices without knowing what faculty members are doing in their courses. This is true even if the scores produced from the person/item interaction within the instrument have valid and reliable properties. Although the revamped end-of-course measure utilized in this study will not solve all problems between students, faculty, and administrators, the psychometric analysis and subsequent findings can help administrators, faculty, and those involved in the dissemination and collection of evaluation data to utilize results in a more responsible way and with a clearer picture of where concerns about results exist.

6. References

American Association of University Professors, Committee C on College and University Teaching, Research, and Publication. *(1974).* Statement on teaching evaluation. *AAUP Bulletin, 60,* 168-170.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Bangert, A. W. (2006). The development of an instrument for assessing online teaching effectiveness. *Journal of Educational Computing Research*, *35*, 227-244.

Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business*, *84*, 40-46.

Beran, T., Violato, C., & Kline, D. (2007). What's the "use" of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, *17*(1), 27-43.

Brown, A., & Green, T. (2003). Showing up to class in pajamas (or less!): The fantasies and realities of on-line professional development courses for teachers. *The Clearing House*, *76*, 148-151.

Calkins, S., & Micari, M. (2010). Less-than-perfect judges: Evaluating student evaluations. *Thought & Action: The NEA Higher Education Journal*, *26*, 7-22.

Campbell, J. P., & Bozeman, W.C. (2008). The value of student ratings: Perceptions of students, teachers, and administrators. *Community College Journal of Research and Practice*, *32*, 13-24.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Retrieved from the IDEA Center website: http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 629-637.

Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education*, *30*, 123–136.

Côté, J. E. & Allahar, A. L. (2007). *Ivory tower blues: A university system in crisis*. Toronto, Ontario, Canada: University of Toronto Press.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*, 292-307.

Ewing, J. K., & Crockford, B. (2008). Changing the culture of expectations: Building a culture of evidence. *The Department Chair*, *18*(3), 23-25.

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, *3*, 1-21.

Franklin, J., & Theall, M. (1989, March). *Who reads ratings: Knowledge, attitude, and practice of users of student ratings of instruction*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Griffin A., & Cook, V. (2009). Acting on evaluation: Twelve tips from a national conference on student evaluations. *Medical Teacher*, *31*, 101-104.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report*. Seattle, WA: University of Washington Press.

Hathorn, L., & Hathorn, J. (2010). Evaluation of online course websites: Is teaching online a tug-of-war? *Journal of Educational Computing Research*, *42*, 197-217.

Hodges, L.C., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. *Innovative Higher Education*, *31*, 279-286.

Jaeger, R. M. (1977). A word about the issue. *Journal of Educational Measurement*, *14*, 73-74.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaplan, R. M. & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications and issues* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. *The Internet and Higher Education, 10*, 89-101.

Kim, K., Liu, S., & Bonk, C. J. (2005). Online MBA students' perceptions of online learning: Benefits, challenges, and suggestions. *The Internet and Higher Education, 8*, 335-344.

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. *National Women's Studies Association Journal*, *19*(3), 87-104.

Linacre, J. M. (2002). What do infit and outfit, mean square and standardized mean*? Rasch Measurement Transactions, 16*, 878.

Linacre, J. M. (2014a). Dimensionality: Contrasts and variances. In *A user's guide to Winsteps Ministep Rasch-model computer programs* (version 3.81.0). Retrieved from http://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2014b). Reliability and separation of measures. In *A user's guide to Winsteps Ministep Rasch-model computer programs* (version 3.81.0). Retrieved from http://www.winsteps.com/winman/reliability.htm

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York, NY: Agathon Press.

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies, 35*, 547–552.

Morgan, D. A., Sneed, J., & Swinney, L. (2003). Are student evaluations a valid measure of teaching effectiveness: Perceptions of accounting faculty members and administrators. *Management Research News*, *26*(7): 17-32.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 87*, 3-15. doi: 10.1002/tl.23

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 86-102). Thousand Oaks, CA: Sage.

Otani, K., Kim, B. J., & Cho, J. (2012). Student evaluation of teaching (SET) in higher education: How to use SET more effectively and efficiently in public affairs education. *Journal of Public Affairs Education*, *18,* 531-544.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis (PCA). *Rasch Measurement Transactions, 19*, 1012.

Sick, J. (2009). Rasch measurement in language education part 3: The family of Rasch models. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *13*(1), 4-10

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation*, *36*, 121–131.

Spooren, P. Brockx, B. & Mortelmans D. (2013). On the validating of student evaluation of teaching: The state of the art. *Review of Educational Research, 83* (4), pp. 598-642.

Toland, M., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272–296.

Thorne, G. L. (1980). Student ratings of instructors: From scores to administrative decisions. *Journal of Higher Education*, *51*, 207-214.

Van Der Ven, A. H. G. S. (1980). *Introduction to scaling*. New York, NY: Wiley.

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

Wattiaux, M. A., Moore, J. A., Rastani, R. R., & Crump, P. M. (2010). Excellence in teaching for promotion and tenure in animal and dairy sciences at doctoral/research universities: A faculty perspective. *Journal of Dairy Science*, *93*, 3365-3376.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques.* Thousand Oaks, CA: Sage.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

Wolfer, T. A., & Johnson, M. M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education*, *39*, 111-121.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.