TRUSTING THE SIMPLICITY OF THE ULTIMATE QUESTION: A CUSTOMER SATISFACTION APPROACH TO STUDENT EVALUATIONS OF TEACHING

Peter Gallo, Creighton University
Cindy Corritore, Creighton University
Chris Wichman, Creighton University
Anne York, Creighton University

Abstract

Student evaluations of teaching (SETs) are important because they often serve as the primary basis for assessing faculty teaching, one of two key inputs into promotion and tenure and annual performance reviews. Our research tests a simple customer satisfaction measure from the marketing literature against traditional teaching evaluation questionnaires. We find that Reichheld's (2003) "Ultimate Question" (UQ) model provides essentially identical results to those obtained from SETs. By not pre-specifying course evaluation questions, the UQ prompts students to describe areas for both process and content improvement. Finally, we describe how electronic content analysis software can use the UQ data to create action-oriented faculty development objectives.

INTRODUCTION

On September 26, 2012, the *New York Times* "room for debate" section posed the question, "Are college students' evaluations of their instructors a useful way to assess professors? What might be more effective?" A number of debaters weighed in on the topic, including "An Insightful Process that Could Be Improved," "Bad Data that Leads to the Wrong Answer," and "As Long as It's about More Than Scores." As expected, most responses came from educators who not only have experienced, but who also dislike, the traditional student evaluation process. However, Jeff Sandefer, founder and master teacher at Acton University, a for-profit entrepreneurship program in Austin, TX, offered, "Our Motto: Give the Customers What They Want." Suggesting that "academics cringe at the word 'customer,'" Sandefer argued that student assessments of teaching are increasingly important and relevant (Sandefer, 2012). "Big changes are coming to higher education, sooner than most people think, and it will include intense competition on both quality and price. Those with business and educational models that don't

listen to their customers are unlikely to survive the tempest -- and frankly, they shouldn't."

Clearly, most educators (as well as the research) agree that traditional teaching evaluations lack validity for a variety of reasons and should not be used in a vacuum, if at all, to assess teaching effectiveness, much less to award tenure, promotions and/or raises. Additionally, the consensus recommendation from several decades of research is that student evaluations of teaching (SETs) should be just one of a number of measures used in evaluating the effectiveness of teachers (d'Apollonia & Abrami, 1997; McKeachie, 1997; Pallet, 2006), though some still argue that any measurement of effectiveness should focus on identifying opportunities for improvement rather than on obtaining a numerical value to rank instructors' teaching (Martin, 1998). However, as the education industry has become more market-driven, accrediting boards such as the Association to Advance Collegiate Schools of Business (AACSB) increasingly look for "evidence-based learning" and increasingly require assessment and improvement of teaching effectiveness. As such, 88% of liberal arts colleges continue to use SETs to evaluate professors' teaching, and 97% of department chairs use them for annual evaluations and tenure and promotion decisions (Berk, 2006).

An answer to the seeming disconnect between research and practice may be found in the final *New York Times* commentators' conclusion: "University administrators widely use student evaluations to assess a professor's teaching quality because they are easy to implement, and there's no well-defined alternative." (Carrell & West, 2012). Traditional evaluations are used as the single teaching-related factor for determining faculty promotion, tenure and compensation because other methods such as observing classroom teaching, reviewing faculty teaching portfolios, and other often-proposed alternatives demand too much faculty and administrative time. That is, they are not "simple" enough.

In undertaking the research described below, we proposed the following question: What if a simpler *and* equivalent method of obtaining student assessments of faculty teaching were available? We attempted to answer this question by exploring an alternative way of obtaining evaluations. Specifically, we compared Fred Reichheld's Ultimate Question (UQ) customer satisfaction methodology (Reichheld, 2003) with a traditional, multi-item SET instrument. Based on data collected from 10 faculty and 478 students, who were asked to respond to both traditional and the Reichheld UQ course evaluation instruments, our findings suggest that results from the simpler and shorter UQ evaluation tool are not only highly correlated with traditional SET questionnaires, but they also have the added advantage of providing students with a quick and open-ended way to recommend the one or two specific changes that would most improve their experience of the course. As such, the Ultimate Question tool may also be more likely than traditional SETs to provide faculty and administrators with concrete ideas for improvement, which can be relatively easily integrated into the annual faculty

review process. We end by describing a process for quickly and electronically highlighting potential areas for teaching improvement using word cloud technology and sentiment content analysis.

SET INSTRUMENT OVERVIEW

A significant body of literature exists on SET instruments, spread across a variety of academic disciplines. While early literature reviews provide a useful distillation of the historical literature on effective teaching behaviors (Feldman, 1976), more recent reviews focus on differences in SET instruments and their uses (Richardson, 2005). Given the breadth of the literature and the availability of well-documented reviews, the overview below will simply highlight the development of SETs' structure and their purpose in an academic institution.

Early literature on effective teaching focused on identifying teacher behaviors that were correlated with effective learning. In Feldman's seminal 1976 article, 19 behaviors were identified, including teacher interest, enthusiasm, subject matter knowledge, breadth of coverage, and organization. This early research led to a multidimensional view of teaching that focused on identifying and measuring instructor behaviors without relating them to a global measure of teaching effectiveness (Marsh, 1987). The multidimensional behavioral perspective has influenced the composition of most SETs used in academia today. In addition, the literature suggests that SETs are generally reliable in the sense of being consistent, stable, and generalizable (Hobson & Talbot, 2001; Richardson, 2005). However, the question remains as to whether SETs are valid indicators of anything other than instructor behaviors (Yunker & Yunker, 2003; Remedios & Lieberman, 2008).

As mentioned in the introduction, the most common use of SETs today is to rate overall instructor teaching and/or overall course quality to provide input for annual performance reviews, as well as for tenure and promotion decisions (Berk, 2006; Sheehan & DuPrey, 1999). This is probably why global measures asking students to rate the overall course quality and the overall quality of a professor's teaching have been increasingly added to the list of SET items. While such global measures may offer a convenient and simple number for those making tenure and promotion decisions and giving performance reviews, Boex (2000) argues that, statistically, the instructor behavior items meant to define teaching dimensions are highly interdependent and, as such, may not correlate well with global measures. Richardson (2005), in a summary literature review, concludes that "students' evaluations are a function of the person teaching the course unit rather than the particular unit being taught." Therefore, SET questions regarding the overall quality of course content or effectiveness of course learning could share the same biases as those regarding teacher quality. Also, because most SETs are designed in-house and have been augmented to include a single global measure of teaching, they are unlikely to have been properly validated (Arreola, 2007). For a more in-depth discussion of SET validity issues, please refer to a series of articles in the

American Psychologist: d'Apollonia & Abrami, 1997; Greenwald & Gillmore, 1997; Marsh & Roche, 1997; McKeachie, 1997.

In summary, while the heavy reliance on these far-from-perfect global measures for assessing faculty teaching for tenure and performance review purposes is a significant cause for concern, in proposing the UQ-based SET instrument, we don't offer a solution to this problem. Rather, realizing that single measures will continue to be used for such purposes, we first offer a simpler way of obtaining them. We then illustrate how what used to be an impossibly time-consuming process of assessing open-ended comments can be transformed into a straightforward tool for developing faculty teaching, thanks to the increasing transition to on-line SETs and the concurrent rise of content analysis software.

IS THERE A SIMPLER ALTERNATIVE TO TRADITIONAL SETS?

Returning to the *New York Times* debate discussed in the introduction, we see a clear movement in the for-profit higher education world toward viewing students as customers (Sandefer, 2012). Some academics and administrators criticize the Academy for being too customer-oriented, while others argue that parents and/or employers are the true customers of higher education (Titus, 2008). Others argue that single-item measures are poor predictors of pretty much everything (Wanous, Reichers, & Hudy, 1997; Thurm, 2006; Pingitore, Morgan, Rego, Gigliotti, & Meyers, 2007). Despite these pushbacks, some support in the SET literature exists for using a single measure to assess teaching effectiveness (Abrami, 1989). Because a parallel argument has been ongoing in the marketing literature for several decades with regard to customer satisfaction, we looked to that research for potential alternatives to traditional SETs (Reichheld, 2003; Morgan & Rego, 2006).

Customer satisfaction literature suggests that organizations first identify specific encounters that may lead to customer satisfaction and then attempt to understand the importance of each activity to the satisfaction rating (Ammar, Moore, & Wright, 2008). Unfortunately, most traditional customer satisfaction instruments cannot effectively identify such encounters, both because they use preset, generic questions to collect data and because they fail to consider which factors most influence customer ratings. Thus, some suggest that traditional customer satisfaction instruments, while relatively inexpensive to design and administer, may yield few actionable results which can be used to identify, track and change behaviors to create a more positive customer experience (York & McCarthy, 2011). Additionally, in their attempt to create reliable and comparable results across business units, firms and industries, and, in some cases, to produce a single number, ranking, rating or average, these instruments have failed to collect the rich information needed to guide improvement efforts.

The lack of a simple method to measure customer satisfaction and the inability to make process improvement changes based on specific areas of concern led Fred Reichheld, a long-time customer loyalty and retention consultant with Bain and Company, to develop his own model, now known as "The Ultimate Question." Reichheld claims that a single question can explain up to 90% of the variance in overall customer satisfaction ratings (Reichheld, 2006). This ultimate question is simple: "On a scale from 0-10, how likely are you to recommend _____ (fill in the blank with product or service) to a colleague, family member, or friend?" The concept behind the UQ is that customer satisfaction and thus customer loyalty is such a strong and value-laden notion that it is usually applied only to one's closest associations – family, friends, or, in a business context, colleagues. Thus, the most natural thing for a customer to do if s/he loves doing business with a particular provider is to recommend that provider to someone s/he cares about (Reichheld, 2006).

From each customer's answer to the Ultimate Question, firms receive a single number, calculated from the UQ 0-10 scale that is used to categorize customers into three groups: promoters (9-10 rating); passives (7-8 rating); and detractors (0-6 rating). While many of Reichheld's proponents have focused only on the Net Promoter Score (promoters minus detractors), in our view the most significant value added of the UQ is the use of a single follow-up question to elicit specific changes that could improve the rating. This additional question extends the single UQ rating from simply providing a satisfaction measure at a point in time to also providing information for future improvement. Many businesses, including GE, American Express and Microsoft have now adopted the UQ customer satisfaction method. This has generated substantial public relations buzz in the marketing, consulting, IT, healthcare, and quality control sectors. However, little published research exists that has empirically investigated the UQ method. This limited work indicates that rather than being superior to existing customer satisfaction methods, UQ performed similarly.

Transitioning from the customer satisfaction research to student evaluation of teaching (SET) research, Palmer and Holt (2009) argue that student satisfaction in the context of the rapidly expanding online teaching environment is especially important to learning. They stated that in this environment, satisfaction is driven by how confident learners feel about their ability to communicate and interact in the online teaching space, in addition to the more traditional clarity of expectations and feedback. To investigate this hypothesis, several recent empirical studies adapted classic marketing service quality satisfaction instruments - primarily SERVQUAL (Parasuraman, Zeithaml, & Berry, 2002) - to assess student satisfaction with teaching (e.g. Stodnick & Rogers, 2008; Kwek, Lau & Tan, 2010; Tuan, 2012). The issues with instruments such as SERVQUAL, while perhaps incorporating more reliable and valid measures than traditional SETs, are that it is lengthy – 19 or so items – and it also, like traditional SETs, uses preset items. Additionally, these instruments seem to do a better job of assessing qualities such

as professor empathy, confidence in a professor's knowledge, and impartiality of assessment (Stodnick & Rogers, 2008) than assessing overall teaching effectiveness.

Even Reichheld (2006) acknowledges that the UQ doesn't work in all industries, especially monopolistic or rapidly growing niches. However, as higher education is neither, the UQ would seem to fit an academic context. We found one paper in the SET literature that explicitly studied Reichheld's UQ model. In their case-based research, Adam and Nel (2009) hypothesize that student satisfaction may be especially important to learning in on-line and blended (vs. traditional classroom) settings. They chose a single item "propensity to recommend" measure rather than a traditional SET due to the lack of correlation with learning outcomes in prior SET literature and the increasing use of single item customer satisfaction measures in industry settings. In the end, they conclude that their single propensity to recommend measure "can only be regarded as a surrogate for other aspects of student learning." (Adam & Nel, 2009)

Clearly, quite a bit of space still exists for empirically exploring not only, simpler methods of assessing teaching effectiveness, but also searching for better ways to develop and improve faculty teaching. The research methodology presented in the following section addresses these two areas. First, we compare Reichheld's UQ method (including a single open ended follow-up question) to a traditional, pre-specified, multi-item SET instrument. Second, we illustrate how improved content analysis and visualization software can provide efficient analysis of students' open-ended responses, thus offering a new tool for both assessment and explicit feedback to improve faculty teaching.

RESEARCH DESIGN

We chose to adapt the UQ model to measure student satisfaction in part because of the parallel nature of the arguments taking place in the marketing literature about using pre-set, multi-item customer service instruments as opposed to a simpler, more open-ended approach. However, because the UQ focuses on respondents' willingness to recommend to friends, the measure seemed uniquely tailored to an industry like higher education, in which indirect cues such as feelings for the professor influence student choices and thus their satisfaction with a particular course. Because the UQ instrument is administered directly at the end of each customer's experience (which, in the case of academia, is around exam time at the end of the course), this survey method had a time frame comparable to tradition SETS, as well as provided an opportunity to ask Passives (7-8 net promoter scores) and Detractors (0-6 net promoter scores) the additional follow-up question: What one or two things could we do to increase your rating to a 9 or 10? As such, the tool had the dual advantage of being able to provide administrators with a single, simple number with which to rank and rate faculty teaching, as well as a method

to hone in on specific student concerns rather than forcing choices into pre-determined categories.

Faculty were asked to volunteer to conduct anonymous Ultimate Question assessments at the same time as the traditional SETs. Nine faculty members, including two of the four authors, participated. Our university's traditional SET instrument contained 19 close-ended, 5 point Likert-type scale items and 4 open-ended questions. Global question 19 (hereafter referred to as T19) of the instrument typically is used by department chairs to determine faculty teaching performance evaluations and raises, as well as tenure and promotion decisions, and reads: "On the whole, the quality of professor's teaching was…."

We modified Reichheld's (2006) UQ in the following way: "On a scale of 0 to 10, with 0 being 'least likely' and 10 being 'most likely,' how likely is it that you would recommend this course to a fellow student and/or friend?" Following Reichheld's (2006) model, the students who rated their likelihood to recommend below 9 (that is, Passives and Detractors) were asked the follow-up question regarding one or two things that could be changed about the course to make them change their rating to a 10. We also asked the Promoters (9-10 raters) to identify one or two things they enjoyed most about the course. The precise methodology for conducting these assessments can be obtained from the authors if desired. Our response rate was 80% (451 students) for the traditional SETs and 66% (415 students) for the UQ instrument.

DATA ANALYSIS AND RESULTS

How Do the UQ Ratings Compare with Traditional Global SET Ratings?

Our first research question was answered by comparing the UQ rating with the SET T19 rating. Both UQ and T19 are scored on Likert-type scales, though the two differ in several ways. The UQ uses a 0 – 10 scale with 10 being "best" (Reichheld makes a point about using the 0 to anchor the scale so that subjects do not get confused about the direction of the scale). T19 uses a 1 – 5 with 1 = Strongly Agree (best) and 5 = Strongly Disagree (worst). UQ responses of 4, 5, or 6 were matched with the traditional SET's neutral value of 3 (labeled Neutral-456).

In the first analysis, a cumulative logistic model was used to explore whether a difference existed between the UQ and T19 ratings. We first ran models for each course/professor combination to account for the correlation between students taught by the same professor. Under the scaled Neutral-456 categorization, 9 of 10 confidence intervals for the odds ratio included 1; thus, generally speaking, there was *no detectable difference between T19 and UQ*. For the second analysis, we used a Wilcoxon's sum rank test (Gibbons & Chakraborti, 2011) to compare the T19 and UQ locations of scores. Prior to conducting Wilcoxon's test, the data were transformed to a level playing field. To accomplish this, we standardized both the UQ and T19 responses by subtracting their respective means and dividing by the

respective standard deviations. Prior to standardization, the T19 responses were reverse coded to make the responses consistent with UQ responses (i.e., the higher the score, the more positive the response). Again, we found *no detectable differences between T19 and UQ* when comparing these standardized scores using the Neutral-456 categorization scheme (see Table 1).

TABLE 1
Summary of Odds Ratio and Wilcoxon Tests

Instructor	Course	Neutral-456			Wilcoxon
		Est. Odds Ratio	Wald 95% Confidence Limits		p-value
1	**BIA253**	0.616	0.271	1.401	0.8481
2	**BUS229**	0.741	0.322	1.705	0.4679
3	**ACC201**	**0.466**	**0.279**	**0.778**	0.1369
4	**BUS471**	1.065	0.492	2.306	0.7698
5	**MKT319**	0.473	0.203	1.098	0.1520
6	**MKT473**	1.047	0.359	3.055	0.8109
7	**ACC202**	1.281	0.520	3.155	0.8748
7	**ACC315**	1.259	0.398	3.986	0.6967
8	**MKT319**	0.422	0.117	1.521	0.3244
8	**MKT363**	0.545	0.154	1.926	0.9132
9	**FIN361**	2.198	0.343	14.102	0.8826

To try to better understand just what the traditional SET items were measuring, we performed an exploratory factor analysis (EFA) on the SET items by professor and by course to determine how many factors emerged, which SET items loaded on those factors, and whether similar factors emerged for each professor. The analysis suggested that the number of factors and items loading on those factors differed widely from professor to professor and course to course. We then conducted a factor analysis on the entire combined SET data set. Factors generated using the minimum proportion of variation explained = 0.75 prior to rotation; factors were then rotated using oblique varimax to allow for possible correlation between factors. The EFA of the aggregation of instructor evaluations is presented in Table 2. The factors identified accounted for 75.22 % of the variation prior to rotation and were identified as follows (in order of importance): Assessments/Grades, Professor Confounded with Course, Workload/Course Difficulty, Professor Availability, In Class Dynamic, and Professor/Student Interaction.

TABLE 2
Factors and Loadings of Traditional Questions for the Aggregate of Classes

Factor	Question Number	Question	Loading (> 0.70)
Assessments / Grades	8	Exams/graded assignments provided a fair, accurate evaluation of my performance.	0.9172
	7	Exams and graded assignments reflected the content and emphasis of the course.	0.8859
	14	The professor's feedback on exams and assignments was valuable.	0.8356
Professor confounded with course	17	In relation to other courses, my overall level of learning in this course was:	0.7835
	10	This course was well-organized.	0.7675
	1	This course was intellectually stimulating.	0.7283
	19	On the whole, the quality of professor's teaching was:	0.7185
Workload / Difficulty	16	In relation to other courses, the workload was:	0.8989
	18	For my academic background and ability, the level of difficulty of this course was:	0.8910
Professor Availability	12	The professor was available for consultation with students outside of class.	0.8978
	11	The professor was responsive when students experienced difficulty w/ material.	0.7106
Class Dynamic	4	Student participation (discussion, expression, questions) was encouraged.	0.8565
	2	The professor was enthusiastic about the subject matter of the course.	0.7862
Professor - Student Interaction	5	The professor was patient and helpful when dealing with students.	0.8531
	11	The professor was responsive when students experienced difficulty w/ material.	0.7307
	3	The professor used class time effectively.	0.6952

Can UQ Methodology Help Develop Faculty Teaching?

To address our second research question, we examined the content of student comments collected from both the traditional SET and UQ instruments using two different semantic analysis tools: Semantria (semantria.com) to analyze the sentiment or tone behind student comment and Wordle (wordle.net) to create word clouds.

Semantria attempts to identify and classify sentiments or tones of text using Natural Language Process (NLP). It uses Wikipedia knowledge to understand and measure distances between words in its analysis. The process identifies facets (nouns or object phrases), attributes (adjectives), entities (proper nouns such as names), sentiment (positive, negative, or neutral), and themes (noun phrases that reflect main ideas in the text) in a body of text. Semantria has been shown to return sentiment scores with a precision rate of 53% (Abbasi, A., Ammar, H, & Dhar, M., 2014). While a more complete description of Semantria is outside the scope of this paper, please see Gabrilovich & Markovitch (2007) and Medhat, Hassan, & Korasy (2014) for additional details.

Sentiment analysis is used primarily to analyze social media text or customer feedback. Thus, because our student comments are a form of feedback, we decided to conduct sentiment analysis on our open-ended student comments. To begin the analysis, we combined all comments across professors and courses and performed two sentiment analyses: 1) focus of comments and associated sentiment, and 2) themes in the comments and associated sentiment. Our results reflected a more granular focus on course details in the traditional SETs as compared with the UQ comments. Essentially, SET comments tended to focus on elements such as tests, time (not enough), quizzes, and assignments. However, the sentiment associated with these items was predominantly neutral. The only negative sentiment associated with detailed course elements was for 'tests'. Students felt they were 'difficult', 'too many', and 'not over the material'. However, this result could be an artifact of the timing of both evaluations (i.e., right before final exams) (Arnold, 2009). The themes identified in the traditional SET comments were also specific and low level. The most common were 'class time', 'group project', 'exam questions', and 'specific software' used in one of the courses. However, these themes all had neutral sentiment. Overall, it appeared that traditional SET comments focused students more on the course details than on a more global course view. It also encouraged a focus on the elements that ultimately were graded (ie. quizzes, assignments, tests). Yet strong sentiment was not apparent. In contrast, UQ students comments tended to address the course overall, with 'class', 'course', 'world', 'material', and 'life' being the most frequently-used terms. 'World' and 'life' were used in the context of 'real world' and 'real life', indicating a focus on applicability of course material outside of the classroom. The overarching themes identified in the UQ student comments mirrored this higher-level focus, reflected by the top themes of 'business world' and 'group work'. Interestingly, the UQ comments also had the same strongly neutral sentiment that was found in the SET comments. In both the UQ and SET comments, however, we noted that the faculty/teacher, while present, were mentioned far less than the course itself.

This focus on the course rather than the professor was also reflected in the Wordle-generated word clouds. Wordle creates a visualization (cloud) of all the words in a body of text in which word prominence is given to words that appear more frequently in the source text. Common words such as 'the' are removed

automatically. Color, position, and orientation are used to convey information. Word clouds have been shown to be useful for preliminary content analysis (McNaught & Lam, 2010) because they quickly identify the main patterns, topics and themes in a text as well as highlight differences between two texts. For example, Clement, Plaisant, and Vuillemot (2008) used word clouds analytically to compare and contrast writing styles in a literary study. We used them in this study to provide a quick comparison of the flavor of the comments from the traditional SET and UQ tools.

The SET word cloud, consistent with the sentiment analysis, reflected a higher frequency of low-level, specific words like 'tests'. For the UQ student comments, in addition to comparing all UQ student comments, we also broke out the UQ comments of all students rating the course either a 9 or 10 (Promoters) from those rating 8 or below (Passives and Detractors), which are the groupings recommended by Reichheld (2006). Using these groupings, we found that the focus we saw in the sentiment analysis on high-level concepts and future usefulness appeared to come from the Promoter group. This is in contrast with the Neutral/Detractor group, which focused on lower-level concepts such as homework, class time, and assignments. Figures 1 and 2 below show the SET and UQ word clouds, along with the passive/detractor vs. promoter comparison groups for the UQ comments.

FIGURE 1
Word Clouds of Student Responses to Traditional vs. UQ SET Instruments

FIGURE 2
Word Clouds of UQ Comments of Passive/Detractor vs. Promoter Students

DISCUSSION AND IMPLEMENTATION

Student evaluations of teaching represent one of the most widely researched yet controversial areas within the science of teaching and learning. We believe this emphasis persists because most universities use SETs primarily to assess faculty performance for tenure and promotion decisions rather than for faculty development. Because of the student time involved in completing the evaluations and the faculty time involved in trying to make sense of them, the primary goal of our work was to explore whether a simpler measure might substitute for the longer traditional SET instrument. We chose to investigate Reichheld's Ultimate Question (UQ) model, which has been widely adopted over the last few years by corporate America to measure and improve customer satisfaction. Reichheld's UQ differs from traditional multi-item SETs in two key ways. First, it asks students to provide a rating for just one question: how likely they would be to recommend the course to fellow students or friends. Second, it asks students one additional open-ended question: to list the one or two concrete areas for course improvement which, in the neutral/detractor group (rating 0-8), would bring the student's rating up to promoter level (9-10).

The results surprised us. When we compared SET question T19 ("on the whole, the quality of professor's teaching was…") with the UQ question, "How likely would you be to recommend…," we found no significant differences as long as we mapped the ultimate question's eleven point scale onto the traditional five point scale by setting 4,5,6 to a neutral value of 3. This finding suggests that students are likely to simply translate the familiar traditional Likert-type 1-5 scale to the UQ 0-10 scale. While we, along with most other faculty and researchers, do not condone the practice of using a single global rating of teaching for faculty performance reviews and tenure and promotion decisions, the pervasiveness and efficiency of this practice suggests that it is not likely to change in the near future. Thus, the comparable but simpler, one-question UQ model is appealing on at least two levels. First, the UQ efficiently provides administrators with the one number they seem to require to assess faculty teaching performance. Second, by not asking for ratings for specific, pre-set faculty behaviors as is done in traditional SET questions, the UQ provides student-generated and more actionable suggestions for faculty development.

To investigate this latter possibility, we employed two content analysis tools to compare the information in students' open-ended comments obtained from both the traditional SET and UQ instruments. The key takeaway was that students completing the SET evaluation's open-ended comments focused much more frequently on low-level assessment-oriented areas of the course such as tests, exams, and graded assignments. Similarly, the factor analysis conducted on the traditional SET 19 Likert-type items also showed that the strongest factors were those covering similar areas. In contrast, students completing the UQ's open-ended question ("one or two suggestions for improvement" or "one or two things

liked best") tended to focus more on broader aspects of the course, such as whether course material reflected the 'real world', 'learning', and 'understanding'. However, interestingly, when the UQ follow-up questions were sub-divided into those rating the course high (Promoters) vs. lower (Passives/Detractors), the high raters turned out to be the ones focused more on global course aspects, while the lower raters' comments more closely resembled the SET comments, focusing on detailed areas such as exams, tests and assignments. One interpretation is that the specificity of the traditional multi-item questionnaire cued students to give narrower open-ended responses in that instrument. However, it also is possible that the timing of the SETs – right before finals – caused grades to weigh more heavily on those students who were unhappy with the course, perhaps because their performance was not as strong as they had hoped.

Implementing our findings turns out to be relatively simple. The UQ instrument could replace the traditional multi-item SET instrument with two questions: one 0-10 rating question and one open-ended follow-up question. Making this change would save faculty, administrators and students time and would probably increase the evaluation response rate. Note that the reason why the rate of our survey response to the UQ was lower than the response rate to the traditional SET is that we were asking students to complete two separate evaluations at the same time. The UQ was unfamiliar and probably seen as repetitive. In future work, we suggest randomly varying the order in which each instrument is completed, which may change this finding.

It stands to reason that asking students to provide just one rating along with one or two open-ended comments would perhaps allow them to provide more thoughtful comments on specific issues that concerned or impressed them about the course. In turn, such comments would be more helpful to faculty and administrators attempting to develop and reward good teaching. Schneider (2013) suggests that using open-ended questions for this purpose would be superior to relying on pre-specified SET items, noting a significant difference in interpretation between students and professors for such items as "The course was well organized." In the case of this question, faculty interpreted the measure as meaning how well course content was tied together, while students assumed that the question referred to how early and often they were reminded of major assignments (Schneider, 2013). Such confusion would not exist with an open-ended question that asks for specific recommendations for improvement; e.g. 'Give us two weeks' reminder before the midterm', 'move lectures on microeconomics to the start of the semester', and 'teach us the methods being used in our industry today'.

The issue with open-ended comments, whether obtained from the UQ or from traditional SET instruments, has always been how to efficiently summarize this feedback in a quick and meaningful way. That is where content analysis software comes into play. Because most teaching evaluations are now on-line, the open-ended comments can be automatically transferred to a file that can then be analyzed

using softwares such as Semantria and Wordle. While setting up these softwares requires faculty to identify up front key descriptive words so that the analysis is meaningful, after the initial set up, the software does the rest of the job, creating easy-to-digest word clouds that can be segmented in variety of ways (such as we did with the above, using detractors vs. promoters, or possibly by gender, ethnicity, GPA, major, class rank, age and so forth). The word clouds could form a more relevant and specific basis for faculty annual performance reviews in terms of suggesting measurable developmental goals and objectives. This sort of analysis could also be used in future research into teaching excellence.

In closing, we return to a caveat mentioned in our paper's introduction. Reichheld and Sasser (1989) established that once customers are no longer captive to companies, they expect total satisfaction the same as customers in competitive markets. We could argue both sides of this question when applying the UQ model in an academic context. On one hand, once a student has selected a university, s/he usually is relatively limited to the professors available at that institution. On the other, assuming that multiple faculty members teach the same course, then there is a choice, even at a single institution.

REFERENCES

Abbasi, A., Hassan, A, & Dhar, M. (2014). Benchmarking twitter sentiment analysis tools. *Proceedings of the International Conference on Language Resources and Evaluation,* Iceland, May 26-31, 823-829.

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education*, *30*, 221-227.

Adam, S., & Nel, D. (2009). Blended and online learning: Student perceptions and performance. *Interactive Technology and Smart Education*, 6(3): 140-155.

Ammar, S., Moore, D., & Wright, R. (2008). Analysing customer satisfaction surveys using a fuzzy rule-based decision support system: Enhancing customer relationship management. *Journal of Database Marketing & Customer Strategy Management,* 15(2): 91-105.

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research*, 48(4): 215-224.

Arreola, R. A. (2007). *Developing a Comprehensive Faculty Evaluation System* (3rd ed.). Bolton, Mass: Anker.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching*. Sterling, VA: Stylus.

Boex, L. F. J. (2000). Attributes of effective economics instructors: An analysis of student evaluations. *Journal of Economic Education*, 31(3): 211.

Carrell, S. E., & West, J. E. (2012). Students confuse grades with long-term learning, *New York Times*. The Opinion Pages.

Clement, T., Plaisant, CL., & Vuillemot, R. (2008). The story of one: Humanity scholarship with visualization and text analysis (Tech Report HCIL-2008-003). College Park, MD: University of Maryland

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52(11): 1198.

Emery, C., Kramer, T. & Tian, R. (2001). Customers vs. products: Adopting an effective approach to business students. *Quality Assurance in Education*, 9 (2): 110-115.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, 5(3): 243-288.

Gabrilovich, E. & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. *Proc. of the International Joint Conferences on Artificial Intelligence*, (7): 1606 – 1611.

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric Statistical Inference,* Fifth edition: 290-299.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52(11): 1209.

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations. *College Teaching*, 49(1): 26.

Kwek, C. L., Lau, T.C, & Tan, H.P. (2010). Education quality process model and its influence on students' perceived service quality. *International Journal of Business and Management,* 5(8): 154-165.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3): 253-388.

Marsh, H. W., & Roche, L. A. (1997). Making student's evaluations of teaching effectiveness effective. *American Psychologist*, 52(11): 1187.

Martin, J. R. (1998). Evaluating faculty based on student opinions: Problems, implications and recommendations from deming's theory of management perspective. *Issues in Accounting Education*, 13(4): 1079-1094.

McKeachie, W. J. (1997. Student ratings. *American Psychologist*, 52(11): 1218.

McNaught, C., & Lam, P. (2010). Using Wordle as a supplementary research tool. *The Qualitative Report*, 15 (3): 630-643.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4): 1093-1113.

Morgan, N. A., & Rego, L. L. (2006). The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Science*, 25(5): 426-439.

Pallet, W. (2006). Abuses and misuses of student ratings. In P. Seldin (Ed.), *Evaluating faculty performance: A practical guide to assessing teaching, research, and service*: 267. Bolton, Mass: Anker.

Palmer, S. , & Holt, D. (2009). Staff and student perceptions of an online learning environment: Difference and development. *Australasian Journal of Educational Technology,* 25(3): 366-381.

Parasuraman, A., Zeithaml, V., & Berry, L. (2002). Servqual: A multiple-item scale for measuring consumer perceptions of service quality. *Retailing: Critical Concepts*, 64(1): 140.

Pingitore, G., Morgan, N., Rego, L., Gigliotti, A., & Meyers, J. (2007). The single-question trap. *Marketing Research,* 19: 9-13.

Reichheld, F. (2006). *The ultimate question*. Harvard Business School Press: Boston, MA.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12): 46-55.

Reichheld, F. F., & Sasser Jr, W. E. (1989). Zero defections: Quality comes to services. *Harvard Business Review*, 68(5): 105-111.

Remedios, R., & Lieberman, D. A. (2008. I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, 34(1): 91-115.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*, 30(4): 387-415.

Sandefer, J. (2012). Universities should use student evaluations to give customers what they want, *New York Times*.

Schneider, G. (2013). Student evaluations, grade inflation and pluralistic teaching: Moving from customer satisfaction to student learning and critical thinking. Paper presented at the Forum for Social Economics.

Sheehan, E. P., & DuPrey, T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology*, 26(3): 188.

Stodnick, M., & Rogers, P. (2008). Using servqual to measure the quality of the classroom experience. *Decision Sciences Journal of Innovative Education*, 6(1): 115-133.

Thurm, S. (2006). One question, and plenty of debate. *The Wall Street Journal,* December 4, B3.

Titus, J.J. (2008). Student ratings in a consumerist academy: Leveraging pedagogical control and authority. *Sociological Perspectives*, 51(2): 397-422.

Tuan, N.M. (2012). Effects of service quality and price fairness on student satisfaction. *International Journal of Business and Social Science*, 3(19): 132-150.

Wanous, J.P., Reichers, A.E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology,* 82: 247-252.

York, A. S., & McCarthy, K. A. (2011). Patient, staff and physician satisfaction: A new model, instrument and their implications. *International Journal of Health Care Quality Assurance*, 24(2): 178-191.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, 78(6): 313-317