Fall 2006 University-wide Course Evaluation Pilot:

Preliminary Assessments of Instrument Functionality, Reliability and Validity

Authors:
Jessica Mislevy and Sharon La Voy
Office of Institutional Research, Planning & Assessment

EXECUTIVE SUMMARY

The University Senate has approved the implementation of a standard, University-wide course evaluation instrument. Each course evaluation will contain a set of universal questions, and may be supplemented by questions from colleges, departments/programs, and/or individual instructors. Across the University, course evaluations will be administered through a web-based system, and students will be able to view the aggregate results to a sub-set of standardized items online.

In Fall 2006, UM began piloting the University-wide course evaluation items. Our goal for this first stage of the pilot was to get a better understanding of how the questions themselves are functioning. The standard items were included on both paper and online course evaluations across four volunteering colleges/departments (ARCH, CLIS, HONR, WMST). Approximately 2100 students completed the University-wide items as a part of their course evaluations for an overall response rate of 66%.

This report addresses the usability, reliability and validity of the proposed University-wide course evaluation items. A series of qualitative and quantitative analyses of the pilot data indicate that:

- When students were asked to comment on items which seemed unclear, were hard to answer, or did not seem to apply, they identified several potential problems. Although the majority of the feedback was positive, a few of the items were classified as non-directional, double-barreled, or not applicable. Concern surrounding the usefulness of information obtained through a series of fixed-choice items alone was another issue raised by the respondents. Students also indicated that the instrument as it stands may be difficult to use for non-traditional courses (i.e., lab or studio courses, graduate courses, courses which are co-taught).

- An examination of the descriptive statistics for the items revealed highly skewed response distributions; that is, the majority of students used only the positive end of the scale (i.e., "Agree" and "Strongly Agree") to rate their courses and instructors. Most of the item means fall around 4.0, with standard deviations around 1.0 scale point. These results, also found at the individual course level, indicate a great deal of agreement among students in their ratings.

- One of the universal items asked students to indicate the amount of effort they put into the course. Response options for this item included: "Little," "Moderate," and "Considerable." When responses to the other 13 items were compared based on reported effort level, the average rating for the little effort group was consistently lower than that of the other two groups, and in most cases, to a statistically significant degree.

- A select group of participants (HONR) completed the standard evaluation form on two separate occasions, allowing us to examine the test-retest reliability of our instrument. Highly correlated scores between the two administrations were obtained, indicating that students respond consistently to the same items at different times.

- We conducted a Principle Components Analysis to explore the dimensionality of our instrument. Our findings advocate a one-component model. All 13 Likert-scale items are highly related to each other and to this one component, suggesting that the standardized questions are targeting a single topic of "overall" course effectiveness or satisfaction. Our results do not seem to indicate that students view items relating to the course and items relating to the instructor as two distinct aspects of course evaluation. Nearly identical results were obtained for both the exploratory and confirmatory portions of our analyses.

- Within this single dimension, we examined the internal consistency among the ratings through Cronbach's alpha. Scale reliability values above .900 for both an exploratory and confirmatory sample indicate that the items "hang together" very well. On average, students' responses remain quite consistent throughout the series of items, providing additional evidence to support the reliability of the ratings.

- We used a measure of split-half reliability to determine whether or not the public items for use by students and the private items for use by faculty and administrators were measuring the same topic of "overall" course effectiveness or satisfaction. We divided the instrument into two subgroups of items and found that the sets of items function almost identically. In essence, for the typical student, the same conclusions regarding his rating of course effectiveness would be drawn from the two subsets of course evaluation items.

Several limitations must be kept in mind while interpreting the results of the pilot. Most importantly, participants were not selected randomly from the population of students at UM. Therefore, results may not be generalizable to all colleges and departments across the campus.

The next phase of our pilot will include periodic review of plans for the system and its results by the University Course Evaluation Advisory Committee. We will also examine the usability and functionality of the technology system – currently in development by OIT – which will ultimately be used to administer University-wide course evaluations. Results from each stage of the system's implementation will be examined and compared to continually assess the stability and consistency of our preliminary findings.

We recommend that the University-wide course evaluations be reviewed even beyond the system's full implementation. The analyses conducted in this pilot should be repeated using a random sample of respondents across the University so that the generalizability of results can be addressed. We also suggest an investigation into the relationship between course evaluation ratings and other indicators of effective teaching to better appraise the validity of the instrument.

BACKGROUND

The University Senate has approved the implementation of a standard, University-wide course evaluation instrument. Each course evaluation will contain a set of universal questions, and may be supplemented by questions from colleges, departments/programs, and/or individual instructors. Across the University, course evaluations will be administered through a web-based system, and students will be able to view the aggregate results to a sub-set of universal items online.

A Senate task force on course evaluation and an implementation committee drafted and finalized items included in the University-wide set, taking into consideration both student and faculty promotion and tenure interests. Because promotion and tenure decisions are personnel decisions by definition, Maryland law requires that information included in the ATP process remains confidential. Therefore, the University-wide items contain a set that students are interested in, and a set that are geared toward promotion and tenure decisions. Students will not have access to the APT items, and APT committees will not have access to the student items. For a list of the Senate-approved universal course evaluation items, please refer to the Appendix of this report.

The system will allow individual colleges to add college-wide items to all of their courses, and the same will be true of departments and programs. In addition, individual instructors will have the opportunity to add items specific to their courses. The instructor will see the results of all of the items (instructor written through University-wide); and college and department administrators will see the department and college items as well as the University-wide APT set. Students who have completed all of their course evaluations the semester before will have access to the results from the University-wide student set the following semester to help guide their course selection. This requirement is waived for new students (i.e., entering freshmen and transfers).

This past fall, UM began piloting the University-wide course evaluation items. Our goal for this first stage of the pilot was to get a better understanding of how the questions themselves are functioning from both a qualitative and quantitative perspective.

PARTICIPANTS AND RESPONSE RATES

Select colleges and departments that already utilized on-line course evaluations through Web-CT were approached to participate in the pilot study of the University-wide items. A college with a wide variety (lecture, studio, etc.) of undergraduate and graduate courses (ARCH), a college of graduate-only programs (CLIS), and two undergraduate programs (WMST and HONR) agreed to participate.

We offered our participating departments the choice of a paper and pencil administration or an online administration through existing Web-CT software. Given that neither of these tools specifically will be utilized once the complete evaluation system is in place, and the mode of administration was not a focus of this pilot, we wanted to be as flexible and accommodating to the needs of our volunteers in this initial stage of the pilot. Our goal was merely to address the functionality of the questions themselves, and not that of the technology used to administer the items.

To further meet the needs of the participants, we allowed the colleges or departments to append additional questions after the University-wide items. Additionally, allowing the participants to add their own discipline-specific items after the standardized questions mimics the design of the future tiered system. We determined that across all participating departments, the University-wide items would appear first, in the same order, and identically worded on every course evaluation.

Table 1 below describes the colleges and/or departments participating in our Fall 2006 pilot.

Table 1. Fall 2006 Pilot Participants and Response Rates

	Evaluation Mode	Selected Sections	Registered Students	Responding Students*	Response Rate
ARCH	Online	67	1323	847	64.0%
CLIS	Online	52	890	639	71.8%
HONR	Paper	20	386	183	47.4%
WMST	Paper	26	566	413	73.0%
Total		**165**	**3165**	**2082**	**65.8%**

* Freshman Connection respondents were excluded as they are not part of our population of interest

It is important to note that no official incentives for participation were provided by UM during this administration. Because we are still in a testing phase with our instrument, the results of these evaluations will not be made public to the students. In the future, offering results to students completing their own evaluations the previous term as an incentive will likely increase our response rates from what we observed in the pilot. We also refrain from statements about the influence of mode on response rates at this time, as neither Web-CT nor paper evaluations will be utilized for the University-wide course evaluations.

QUALITATIVE ANALYSIS

We asked students participating in the pilot to comment specifically on items which seemed unclear, were hard to answer, or did not seem to apply to them. Of the students responding to the open-ended item, a notable proportion generally stated that the questions "looked good" or "seemed fine." Another large group simply wrote a response such as "N/A," "none," or "no comment." Overall, the bulk of students indicated the questions were reasonably clear and easy to answer. Several students even described the questions as "self-explanatory." One student exemplified this view when he or she stated, *"These questions were the most direct and specific questions I have ever seen on a course evaluation. The questions got to the root of issues that often arise in class situations."*

Other students, however, felt the items were too vague, broad, or generic. They were concerned that the information obtained from the fixed-choice items may not be very useful in terms of providing the instructor with specific feedback or making organizational changes. As a result, many students expressed a desire for additional comment boxes and/or text boxes after each item. They felt they could not convey the complexities of their opinions and experiences through the closed-ended response options alone.

One pattern emerging from the comments related to the appropriateness of a standardized form for all students and courses. A group of graduate students indicated that the University-wide items seemed geared towards undergraduate courses. Along the same lines, students in studio courses found the items to be less applicable to their classroom experiences. Several respondents said that they felt the form was difficult to complete for both their TA and professor; they indicated that their graduate assistants and professors have such different tasks, and that they were not always aware of who was responsible for certain aspects of the course. The frustration and confusion felt by these students was represented by one student who indicated that *"This course was co-taught. Some of my answers about teaching style and* *preparedness apply to one teacher but not the other. Adjusting the evaluation to allow for multiple teachers may be necessary."*

Several students noted that a few of the University-wide items were non-directional. Item #4 ("The instructor set appropriately high standards for students.") and Item #12 ("The workload was appropriate for the course level and number of credits.") seemed to be problematic for this reason. For example, in reference to Item #4, one student stated *"…disagreeing can mean that the teacher set standards too high or too low. It doesn't specify which one."* Similarly, students commented that they were unable to indicate whether the workload was too heavy or too light through the response options supplied for Item #12.

A number of students stated that Item #6 ("Overall, this instructor was an effective teacher.") and Item #7 ("The instructor was effective in communicating the content of this course.") were redundant. They were unable to determine the difference between these seemingly repetitive items.

Several students also suggested that a few of the University-wide items were double-barreled. A double-barreled item is problematic because the item conveys two or more ideas, and endorsement of the item might refer to either or both ideas.[1] Item #8 ("Course expectations and guidelines were clearly explained at the beginning of the semester.") was one item explicitly classified by respondents as double-barreled. For instance, one student commented *"Question 8: Expectations and guidelines are different and should be treated as separate questions."* Other respondents seemed to implicitly classify Item #2 ("Course materials were well-prepared.") as double-barreled. In reference to Item #2, one student stated *"It might be more helpful to separate questions. For example, in this case the professor's lectures were well-prepared, organized and presented well. However, the total opposite was true for the assignments and tests. Essentially evaluating the professor's teaching style, lecture/presentation, assignments, and exams are all separate entities and may be evaluated differently."* Several students indicated that they would like to see distinct questions for various types of course materials.

Item #11 ("The grading in this course was fair.") was often classified as "not applicable" in the comments from students. A number of students stated that they could not provide a valid answer to this item because they had not received their final grade. Some students said they had only received feedback on their performance for a few assignments, and that they were not able to form an accurate opinion based on this limited information. For instance, one student replied *"It is difficult to evaluate the answer to Question 11, as we have not received our final grades yet. I do not have enough graded assignments from the class to conclude anything about the grading."* Another respondent had difficulty selecting an appropriate answer to this item because he or she was unaware of the grades received by other students in the course; this respondent found it awkward to rate the fairness of the instructor's grading based on personal experience alone. A few students also suggested that this item would be improved with alternative response options. These students would have preferred answer choices such as "too easy," "fair," and "too harsh" which would allow them to provide what they viewed to be more useful information.

Several students felt that answering Item #13 ("I would recommend this course to other students.") was a futile effort. The students expressing this point of view appeared to be evaluating courses which were requirements for their major or program. Thus, they did not see the point of recommending – or failing to recommend – a course to their fellow students. For example, one respondent stated *"[Item #13] seems to be a useless question for this course being that every student in this major has to take this course."*

[1] Robert F. DeVellis, *Scale Development: Theory and Applications* (Thousand Oaks: Sage Publications, Inc., 2003): 68.

Respondents highlighted a few aspects of their courses which they felt were not covered by the evaluation questions. A number of students indicated that they would like to see an item related to the course textbook and/or reading materials. Occasionally, respondents stated they would also be interested in a question about appropriate and timely feedback from instructors on completed assignments. Finally, several students would like an open-ended item asking for recommendations and suggestions specifically, in addition to general course comments.

Respondents appeared to have mixed feelings regarding the length of the course evaluation form. Some students said they were pleased to receive such a short and concise form, while others said the number of items was excessive. Unfortunately, it is not possible to determine the students' opinions on the length of the University-wide assessment alone since each student was also asked to answer the additional questions appended to their evaluation as requested by their department. The number of supplementary items was also dependent on the type of course the student was evaluating (i.e., lecture courses in one department may have received a different number of items than a studio course in the same department). It should be noted, however, that fewer students receiving evaluations with approximately 20 items seemed to be dissatisfied with the length than students receiving evaluations with upwards of 40 items.

Respondents also had comments on the Web-CT system used to administer the online course evaluations. Although Web-CT technology will not be utilized for the University-wide course evaluations, it may be important to keep the students' comments in mind as the new technology system is developed. In general, for those completing a web evaluation, respondents appreciated the ability to complete their evaluations online. They said the online evaluations provided them with increased privacy and greater flexibility. A significant number of students were frustrated that they had to save their answer to each question before they were allowed to submit their evaluation. In order to make the form more efficient, they requested a "save all" option. A few students also stated that the continual email reminders asking them to complete their evaluations became unnecessary and irritating. Please note that the new technology system which is currently being developed will not require students to save their answers to each individual item.

Concerns surrounding anonymity and confidentiality were also raised by several respondents. These students felt their answers could easily be linked to their name, as they were asked to log into the Web-CT system using their directory ID and they continued to receive email reminders if they had not completed their evaluations. Demonstrating this view, one respondent stated *"My concern is that the evaluations are not anonymous since we log into Web-CT with our user profiles. I would imagine some students are hesitant to evaluate honestly because of confidentiality concerns prior to the completion of the semester and final grades."* One student also felt uneasy because his or her name was displayed across the top of the computer screen while completing an evaluation. Note that the online evaluation form indicated that student responses would be treated as confidential, not anonymous. Additionally, the directions on both forms stated that results of the evaluation would not be released to instructors until final grades had been submitted.

QUANTITATIVE ANALYSIS

We begin the quantitative examination of the evaluation instrument by reviewing descriptive statistics associated with each of the items across all 2082 respondents. See Table 2 below. Note that a small proportion of "not applicable" and blank responses have been excluded. Valid percents rounded to the nearest whole number are presented in the table.

Table 2. Descriptive Statistics across the Entire Sample (n ≈ 2082)

	% Strongly Disagree	% Disagree	% Neutral	% Agree	% Strongly Agree	Mean	Stdv.
1. The instructor treated students with respect.	1	2	4	28	65	4.55	.733
2. Course materials were well-prepared.	2	5	11	34	48	4.22	.953
3. The course was intellectually challenging.	2	4	12	38	44	4.19	.914
4. The instructor set appropriately high standards for students.	1	5	13	44	36	4.08	.912
5. I learned a lot from this course.	2	4	11	36	47	4.22	.931
6. Overall, this instructor was an effective teacher.	3	5	12	35	46	4.16	.992
7. The instructor was effective in communicating the content of the course.	2	6	12	37	44	4.13	.983
8. Course expectations and guidelines were clearly explained at the beginning of the semester.	2	5	11	39	44	4.16	.945
9. The instructor was responsive to student concerns.	2	3	10	34	52	4.32	.876
10. The instructor helped create an atmosphere that kept me engaged in course content.	3	8	15	34	41	4.03	1.053
11. The grading in this course was fair.	3	5	14	38	41	4.10	.976
12. The workload was appropriate for the course level and number of credits.	3	9	12	42	35	3.95	1.051
13. I would recommend this course to other students.	4	7	15	32	41	3.99	1.112

For every item, both the minimum and maximum scale values were utilized by at least small a proportion of respondents. Typically, less than 1% of all respondents selected the "not applicable" response option accompanying an item. For two items (Item #11 and Item #13), however, this figure jumped to approximately 3% selecting "not applicable." This discrepancy may be explained in part by the discussion surrounding these items in the qualitative section above; several students indicated that they could not evaluate the fairness of the grading before receiving final grades, and others stated that it did not matter whether or not they would recommend a course since the course was a program requirement.

With the arithmetic mean for each item hovering slightly above or below 4.0, we see that the distributions for these items are all negatively skewed; that is, most students appear to have used the positive end of the scale (i.e., "Agree" and "Strongly Agree") to rate their courses and instructors. Skewness statistics for the 13 items range from -2.085 to -1.01 with corresponding standard errors around 0.05, further confirming this observation. The standard deviations from the mean for each item all fall around 1.0 scale point. Traditionally, researchers prefer items which show greater variability in responses for analytic purposes. In the special case of course evaluations, however, a lower variance estimate indicates a fair amount of agreement among students. Thus, the similar ratings concentrated towards the upper-end of the scale suggest consistency in the students' attitudes. Although the results from individual courses are not addressed in this document, similarly skewed distributions with means near 4.0 and standard deviations around 1.0 scale point were obtained for the majority of course-level evaluations.

We also hoped to show that our results support intuitive and anecdotal findings related to amount of effort put into a course and satisfaction with the course; that is, we anticipated more positive ratings from students who indicated that they put higher levels of effort into the course. To examine expected patterns in our data, we used the remaining scaled course evaluation item ("How much effort did you put into the course?") to compare the responses from students who indicated that they put "Little," "Moderate," and "Considerable" amounts of effort into the course. Overall, almost two-thirds of respondents (61%) indicated that they put "considerable" effort into the course, whereas 36% indicated a "moderate" level of effort, and 3% indicated that they put "little" effort into the course. Using a robust test for the equality of means, the differences in the average response between the three groups are statistically significant at the 0.05 level for all but two of the items (Item #8 and Item #9). The magnitude of the group differences tends to range between 1/2 and 1 scale point across the 13 items. For almost all of the items, respondents indicating that they put "little" effort into the course provided lower ratings on average than students indicating a "moderate" or "considerable" level of effort.

A visual inspection of the response patters, however, reveals this trend is not linear across all of the items; although the "little" effort group is consistently lower than the other two groups, at times, the mean of the "moderate" group is approximately equal to or higher than the "considerable" group. The only item for which this statement does not hold true was Item #11 ("The grading in this course was fair.") For this item, the group indicating that they put a "considerable" amount of effort in the course has the lowest average rating of all the groups, although this difference of 1/5th of a scale point may be of little practical significance.

In addition to an examination of descriptive statistics for the items, we wished to ensure our instrument is psychometrically sound. As one part of this process, we assessed the stability of ratings over time through an examination of test-retest reliability indices. If obtained, highly correlated scores between the two evaluations, using the individual student as the unit of analysis, would help to provide evidence that the evaluation yields stable and reliable results. In order to calculate the test-retest reliability of scores from the instrument, we administered the form twice to the same set of students with permission from one of our participating departments (HONR). We requested that instructors allow for a one-week interval between the two administrations. Along with two sets of paper forms, instructors were provided with a separate set of instructions for each administration. Participants were not informed of the second administration at the time of the first administration, as we wished to limit possible influences on participant motivation and testwiseness.

Seventy-seven students across seven Honor's seminar courses completed the same evaluation form on two separate occasions. Approximately 400 HONR students within 20 course sections had been selected to participate in the test-retest portion of the pilot. Unfortunately, more than half of these sections were unable to administer the form twice for various reasons.[2] As a result, we elected to use the individual as our unit of analysis, as opposed to the course. We also chose to examine the correlation between the Time 1 and Time 2 responses for each item because the number of cases limited the accuracy of more advanced reliability indices. Note that the descriptive statistics of the HONR responses to the items closely mirror those of the entire respondent pool. To best represent the nature of our data, a Spearman's rho statistic correlating scores from Time 1 and Time 2 was calculated for each item. The obtained values range from 0.506 to 0.828, and all correlations are statistically significant at the 0.01 level (2-tailed). These coefficients of stability provide evidence to suggest that students respond consistently to the same items at different times.

For the remaining analyses, the respondent pool was randomly divided into two groups of approximately equal size: an exploratory sample and a confirmatory sample. We utilized the exploratory sample as our primary development sample and the confirmatory sample to replicate and cross-check our findings.[3] When estimates remain fairly consistent across the two samples, it is unlikely that these values are distorted by chance. Thus, the estimates from the two sub-samples provide valuable information about scale stability. It is important to note that sub-samples created from one entire sample are expected to be more similar than two completely different samples taken at two different time points. For our purposes, the same analytic techniques were utilized on the items for both samples, first to determine the underlying structure of the data, and second to confirm or contradict a particular pattern of relationships predicted on the basis of our exploratory results.[4]

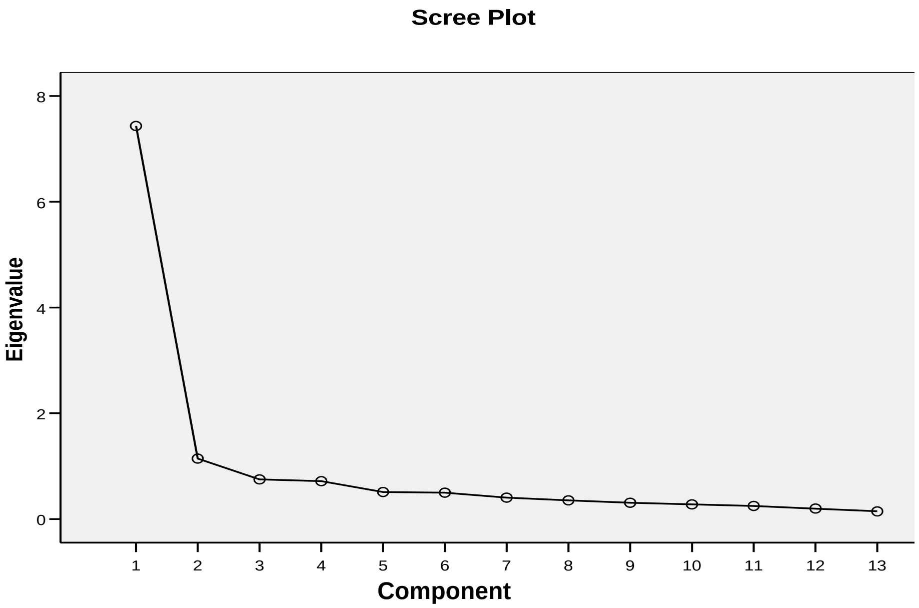
To continue our investigation regarding the reliability and validity of our measure, we conducted an exploratory factor analysis utilizing the 13 Likert-scale items. Through this analysis, we hoped to determine empirically how many constructs underlie the set of items. Or, alternately, we wished to conclude whether one broad or several more specific components were needed to characterize the item set.[5] We selected a principle components analysis method of extraction in order to accommodate the multicollinearity among our highly-correlated variables.[6] Our sample size – even when split for the exploratory and confirmatory analyses – is more than adequate for a PCA with well over 50 respondents per item.[7] To further ascertain the factorability of our correlation matrix, we utilized Bartlet's Test of Sphericity ($p_{exploratory} = p_{confirmatory} < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy ($KMO_{exploratory}=0.931$, $KMO_{confirmatory} =0.938$)[8]; the results of both tests suggest that our matrix is factorable.

We ran a principle components analysis on the 13 items for our exploratory sample. We then utilized several popular guidelines to determine the number of factors underlying the data.[9] In one approach, factors with an eigenvalue greater than 1.00 are retained. In our analysis, one factor had an eigenvalue notably over 1.00 ($\lambda_1=7.432$), and a second factor with an eigenvalue hovering just above 1.0 ($\lambda_2=1.143$).

[2] Of the 20 selected HONR seminars, seven sections completed both forms, eight sections completed only one form, and five sections completed neither form. See the limitations section of this report for details.
[3] DeVellis 99.
[4] DeVellis 131.
[5] DeVellis 103.
[6] For both the exploratory and confirmatory sample, the determinant of the correlation matrix was near zero.
[7] A total of 1292 students answered all 13 Likert-scale items. Blank and "not applicable" responses were removed from these analyses.
[8] Marjorie A. Pett, Nancy R. Lackey, John J. Sullivan, *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research* (Thousand Oaks: Sage Publications, Inc., 2003): 72-78.
[9] Pett 115.

A visual examination of the scree plot suggests a one-factor model, as there is a distinct drop in amount of information (i.e., eigenvalue magnitude) across the successive factors.[10] See Chart 1 below.

Chart 1. Plot of Extracted Factors against their Eigenvalues in Descending Order of Magnitude



It appears that one component does an adequate job of accounting for covariation among the items. This single component accounts for 57.2% of the total variance, and any additional factors would each account for only a small portion of the variance (<10%). Because the second component explains less than 10% of the remaining variance, utilizing one factor to explain a 13-item scale is a parsimonious solution. The extracted loadings on the first factor (i.e., correlations between the items and the factor) range from 0.591 to .883 for the 13 items, with the majority of loadings above 0.700. This finding suggests not only that the individual items are heavily related to this one component, but also that the component is reliable. In addition, the presence of a single factor suggests that it is possible to compute only one component score for each respondent – as opposed to 13 individual item scores – to reasonably describe how the respondent answered the series of items.[11]

We ran an identical PCA technique using the data from the confirmatory portion of our sample and found strikingly similar results. In this analysis, eigenvalues of 7.528 and 1.073 associated with the first component and second component, respectively, were obtained. The single-component model accounts for 57.9% of the total variance in this sub-sample. Again, the item loadings on the first factor range from 0.635 to 0.882. The presence of all items loading heavily on the same factor for both the exploratory and confirmatory portions of our sample provides evidence of construct validity. Our results do not seem to indicate that students view items relating to the course and items relating to the instructor as two distinct aspects of course evaluation. The series of standardized evaluation questions appear to be targeting a single topic of “overall” course effectiveness or satisfaction.

[10] DeVellis 114.
[11] DeVellis 103.

Within this single dimension, we examined the internal consistency among the ratings through Cronbach's alpha. This measure of internal consistency is based on the average correlation among items. For our exploratory sample, the 13-item scale had a Cronbach's alpha of 0.936. Removing any of the items from the scale would not improve the alpha statistic by more than 0.001 units. An alpha of 0.938 was obtained for the confirmatory sample. Again, deleting any of the items would not improve the internal consistency of the 13-item scale. With alpha values falling well above an acceptable threshold of 0.700 for an adequate scale, we have additional evidence suggesting the reliability of the ratings. On average, students' responses remain quite consistent throughout the series of items.

Lastly, we explored the split-half reliability of the course evaluation instrument. This analysis was conducted to determine whether or not the public items for use by students and the private items for use by faculty and administrators were functioning as two parallel subtests. If obtained, high split-half reliability would further suggest that the two subsets of items are measuring the same construct. We were able to calculate the coefficient of equivalence, or the parallel forms reliability, for these two sets of items. For the exploratory sample, the obtained Spearman-Brown split-half coefficient for the two subscales of unequal length is 0.906. Similarly, for the confirmatory sample, a reliability coefficient of 0.910 was obtained. Typically, a coefficient above 0.700, and preferably above 0.800, indicates an acceptable level of reliability. Both of the reliability coefficients obtained in our analyses surpass the standard criteria. In essence, for the typical student, the same conclusions regarding his rating of course effectiveness would be drawn from the two subsets of course evaluation items.

LIMITATIONS

Several limitations must be taken into consideration while interpreting the results of the analyses. First, our pilot utilized a non-probability sampling design; that is, elements were selected on the basis of their availability to form what is known as a convenience sample. As a result, certain members of the student population at UM had no chance of being included in our investigation, and it cannot be determined the extent to which our sample actually represents the entire population.

Second, to accommodate the needs of our participants, we utilized a mixed-mode design. For this pilot, we administered both an online and paper version of our survey, and combined the results across the two modes. As a result, the potential for measurement error increased because the mode may have had an impact on the way respondents answered the questions, even though the order and wording remained identical. We are assuming that the effect of using multiple modes is benign, and that our instrument is relatively immune to changes in the data collection mode.

Third, along these same lines, students across the four participating colleges/departments were grouped together for the quantitative analyses. The clustered nature of the data was ignored for these exploratory analyses, as the clusters (i.e., participating departments and/or courses) were not selected randomly from the population. Therefore, it is not possible to correct for potential differences across the groups in a way which would allow us to generalize our findings to the entire campus community. We are unable to model the sampling error for our design, so attempting to adjust for between-group differences would be virtually meaningless.

Fourth, the results from the test-retest reliability portion of this pilot specifically may not be highly generalizable. A number of instructors never retrieved the evaluation forms after their distribution, so we were unable to obtain either Time 1 or Time 2 data for these participants. Additionally, the response rate for sections participating in both administrations was low; only 7 out of the intended 20 seminar courses participated a second time. As a result of these complications, our intended sample size of approximately 400 was drastically reduced, and the data we were able to collect may be subject to nonresponse bias.

Finally, the length of time between the two administrations was not consistent across the HONR courses. Instructors had been asked to administer the forms exactly one week apart, however, due to time restrictions and their lesson plans, instructors administered the second evaluation after a variable length of time. It is possible that respondent motivation and testwiseness had more or less of an impact on results depending on the length of time between the two administrations.

CONCLUSIONS AND FUTURE DIRECTIONS

At this point in time, our assessments of the standardized course evaluation items have been primarily exploratory in nature for the purpose of hypothesis building and testing. We intended to provide insights solely related to the items included on our instrument, and not the administration mode. Overall, the results from this pilot suggest student responses are quite consistent across the items. The questions seem to be highly related, targeting a single aspect of "overall" course effectiveness or satisfaction. Respondents, however, have identified several potential problems associated with the wording of select items.

IRPA is coordinating an advisory committee which will periodically review the plans for the course evaluation system and its results, and comment on the decisions that will have to be made as the system is developed. The University Course Evaluation Advisory Committee is comprised of individuals across the campus nominated by the dean of their college as a representative, along with original members of the Senate Task Force that investigated the course evaluation proposal. The student feedback will be discussed at an upcoming meeting of the committee, along with the quantitative results. Any potential changes to the item wording will be addressed at that time.

An on-going review at each stage of implementation is strongly recommended. Although we split our sample into exploratory and confirmatory subsets, to fully gauge the psychometric properties of our instrument, we must replicate our analyses over time and across different samples. If the item set continues to perform as our preliminary analyses suggest, we will have stronger evidence to support the reliability and validity of our findings.

In the second stage of our pilot, we will also investigate the effectiveness of the technology system which will be used to administer course evaluations. It will be important to determine the stability of our findings across this alternative mode of administration. Additionally, we will examine the impact of any changes in the wording of our items.

Further, we recommend that the University-wide course evaluations are continually reviewed beyond the system's full implementation. Eventually it will be possible to take a random sample of colleges, departments, courses and/or students from across the University. The analyses conducted in this pilot should be replicated using this probability sample so that the generalizability of results may be evaluated. The response rates should also be monitored and addressed to help reduce the potential for nonresponse bias and ensure results are representative.

Finally, we suggest supplementing the pilot analyses with additional investigations focused on the instrument's reliability and validity. As a portion of this pilot, we were able to take a preliminary look at the short-interval test-retest reliability of ratings; it may also be beneficial to compare current end-of-course ratings to ratings of the same course and instructor the following year. Once the system is fully implemented, it will also be possible to link responses back to students for research purposes. The ratings of graduate students versus undergraduates, or majors versus non-majors, etc., could be compared to identify patterns. To better appraise the validity of the instrument, we recommend the relationship between course evaluation ratings and other indicators of effective teaching be explored.

APPENDIX

Proposed Universal Questions for Course Evaluation Instrument
Draft April 13, 2006

Questions 1-13 answered on the following scale:
Strongly Disagree Disagree Neutral Agree Strongly Agree (Not Applicable)

Questions for use by faculty and for administrative purposes only (developed by faculty in reference to what they consider in APT and other decisions)

1. The instructor treated students with respect.
2. Course materials were well-prepared.
3. The course was intellectually challenging.
4. The instructor set appropriately high standards for students.
5. I learned a lot from this course.
6. Overall, this instructor was an effective teacher.

Questions for use by students (developed by students in reference to what they would like to know in order to choose courses)

7. The instructor was effective in communicating the content of the course.
8. Course expectations and guidelines were clearly explained at the beginning of the semester.
9. The instructor was responsive to student concerns.
10. The instructor helped create an atmosphere that kept me engaged in course content.
11. The grading in this course was fair.
12. The workload was appropriate for the course level and number of credits.
13. I would recommend this course to other students.
14. How much effort did you put into the course? (Little, Moderate, Considerable)

Open text prompt (responses available only to instructor or department)

15. Additional comments (e.g., about course content/materials, teaching style, etc.):

Very last question (after college/department/instructor)

16. How does this course fit into your academic plan or course of study? (CORE Requirement, Major/Certificate/Minor/Program Requirement, Elective)