MEMO to UIS Faculty on the Summative Teaching Evaluation Report March 20, 2007

Dear Colleagues,

A few years ago, the senate decided to study the instrument UIS uses to measure student perceptions of teaching. Because student perception instruments have two very different and often conflicting purposes--formative to assist faculty in improving their teaching, and summative to help personnel committees make decisions--we eventually divided the work among two separate ad hoc committees which we cleverly called the "formative" and "summative" groups. We asked the summative group to dig into the years of UIS data, to study the validity and reliability of the instrument and to look at its usage by personnel committees. The report of the "summative" group (Professor Lynn Pardie, and Assistant Professors Eric Hadley-Ives and Joseph Huff) is enclosed.

I'd like to give you a bit of background about the UIS instrument. A number of us have railed against it for years (decades, in my case). On the formative side, many of us decided it was utterly useless in helping us to improve our teaching. Many of us adopted a supplemental instrument (an addition to the official instrument each semester), as well as mid term instruments. I think there is consensus that it is not a sophisticated formative instrument. I know that the "formative" ad hoc group will soon have a set of recommendations that would allow each faculty member to individually select additional questions that they want their students to answer, with the results reported only to that faculty member. The campus also desperately needs a unit which actively assists faculty in becoming excellent teachers, and which should include opportunities for formative peer reviews of teaching.

On the summative side, many of us assumed the UIS instrument was not valid and that getting rid of it would be a great service. The evidence and analysis in the report challenges those assumptions and a number of other widely held faculty beliefs. To put it in simplistic, non-technical terms, the report concludes that the instrument is generally valid and reliable, and that the instrument is good at making gross distinctions: who is in the ball park (which most of us at UIS are) and who is out of the ball park. The report suggests we are not using the data legitimately when we try to make fine distinctions and comparisons between those of us who are already inside the ball park, and that personnel committees need training in how to use the data. Among toher things, it recommends that we use the appropriate data set when we judge faculty, as in general, evaluations tend to be lower in lower division, better in the upper division, and better yet in graduate studies.

Please take a careful look at this report. Whatever conclusions we draw from it, I hope that we can collectively tone down our rhetoric about the merits of the instrument until we have taken a hard look at the evidence and analysis and had a thorough discussion. (I will be first in line to say I have stopped my railing.)

The senate will begin its discussion of the report on Friday and continue it over a few meetings. We will consider the recommendations in the report, including the development of guidelines on the appropriate use of the data by personnel committees, as well as systematic training.

If you have any questions or comments, please attend the senate meeting or contact me. --Pat

Background and Charge to the Summative Teaching Evaluation Committee

An Ad Hoc Committee on the Evaluation of Teaching submitted a report to the UIS Campus Senate in the Fall Semester of 2004, in which they questioned the soundness of the rating form that is currently used at UIS to collect student perceptions of teaching. Although the Ad Hoc Committee's report addressed broader issues regarding the evaluation of teaching through portfolio methods and in relation to online versus on-ground courses, it also detailed several concerns specific to the student rating form and its use in the personnel process. For the most part, the Ad Hoc Committee's concerns about the form were focused on the following issues:

- the summative nature of the form
- the validity of some items
- the impact of administration guidelines and response rates
- the potential for student biases to differentially affect results
- the lack of uniform personnel guidelines for interpretation of aggregated results

In January of 2006, the UIS Campus Senate asked Lynn Pardie (Professor, Psychology Program), Eric Hadley-Ives (Assistant Professor, Liberal Studies Program), Joseph Huff (Assistant Professor, Management) to follow-up on the Ad Hoc Committee's concerns by assessing the psychometric quality of the current UIS rating form, and to make recommendations regarding the viability of the current form for continued use as a summative measure.

Scope of the Work Conducted by the Summative Teaching Evaluation Committee

In preparation for the Committee's work and this report, the Committee members (a) reviewed relevant academic literature concerning the development and use of student ratings of teaching, (b) attended an interactive audio conference (*Student Ratings: Their Design, Construction, & Use*) conducted by a nationally recognized expert in the field of instructional evaluation, (c) reviewed the technical documents describing the development of UIS' current student rating form, and (d) worked with the UIS Associate Provost for Information Technology, Farokh Eslahi, to conduct statistical analyses[1] of archival data and assess the psychometric quality of the current student rating form. Our goals were to compare the descriptive statistics for the current and original normative data; to determine whether there had been any significant fluctuations in the mean scale scores on critical items; and to assess the extent to which item validity might be compromised by non-teaching related extraneous factors such as gender and expected grade. The remainder of this report summarizes the Committee's findings and recommendations.

Psychometric Characteristics of the Current UIS Student Rating Form

Item Development and Original Psychometric Characteristics

Student rating forms can be developed to serve different purposes. Forms developed for formative evaluation purposes typically consist of many specific items addressing a wide variety of course-related and instructional features. Formative evaluation items are designed to provide the instructor or faculty member with student perceptions that might prove useful in improving teaching effectiveness. Summative forms typically have fewer items, and the items are more global or summary in nature. According to Cashin (1999):

> Global student rating items tend to correlate more highly with student learning than do more specific items….Using a form with only a few items has definite advantages. Such items apply to a wide variety of courses (probably to all courses), and so could be used as the basis of comparison across the institution, as long as the appropriate comparative data was available. (pp. 34-35)

Summative forms are constructed to provide a measure of the overall quality of the teaching as perceived by students. Formative evaluation serves a faculty development purpose, while summative evaluation can provide the basis for comparisons across faculty. Items on the UIS rating form were originally developed to serve a summative purpose.

Several of the items on the UIS form are demographic or categorical in nature. The items that are a focus in the personnel process are the two constructed using a continuum response format and allow for comparisons with normative data across all courses at UIS. The two critical items on the current version of the student ratings of teaching form were originally developed in the early 1970s by a statistician at UIS. The original items differed slightly in wording from the present item forms. The original items were written as:

#8. *Do you think this teacher is competent in the content or matter offered in this course?*

exceptionally competent		satisfactory		incompetent
5	4	3	2	1

#10. *Overall, do you consider this person a good teacher?*

excellent		good		poor
5	4	3	2	1

Colliver, the statistician who originally developed the items, documented the careful and professionally sound approach he used to evaluate the psychometric validity of the items in a series of technical reports (Colliver, 1972; Colliver & Wesley, 1976a & 1976b). The two items were strongly and positively related; as student perceptions of teacher competence increase, ratings of overall quality also increase (.79 to .80 over two academic quarters).

Reliability of course means was assessed by averaging the item ratings for each course and faculty member, and then correlating ratings across a sample of courses, and then across consecutive semesters for all faculty members. Results were statistically significant and indicated moderate-to-high levels of consistency in ratings across different courses and quarters. Correlation coefficients for the sample of courses ranged from .36 to .65, and all were statistically significant. The correlation between the mean spring and winter quarter ratings for each faculty member was .62. The reliability of the quarter means was higher because it was based on more course-related information. Although the items had only been used for two semesters, Colliver estimated that the reliability of a grand mean calculated from all courses taught over three consecutive semesters would be .83.

Colliver also reported an estimated standard error of measurement for grand means (i.e., the average of available ratings for each faculty member) of 0.13. According to Colliver, “it can be said that the hypothetical ‘true’ grand mean for any faculty member will be included within an interval defined by the faculty member’s obtained grand mean plus or minus .13 about 68% of the time” (Colliver, 1972, p. 14). Colliver evaluated the range of the grand means and found it to be 1.5 (3.5 to 5.0), which meant the standard error of measurement was less than 1/10 of the range – well within acceptable limits. Colliver also calculated the standard error of the mean for the initial group of 75 faculty members. Over 75% of the faculty had standard errors of the mean that were smaller than the standard error of measurement for the grand mean (i.e., less than .13).

Although all of the means fell above the midpoint of the rating scale, Colliver wrote, “The literature on faculty evaluations…shows that this is to be expected…When rating faculty, students have a tendency to almost exclusively use the upper end of the rating scale” (1972, p. 15).

After determining the reliability of the items, Colliver sought to assess their validity – the extent to which the items actually measured student perceptions of teaching quality. He recognized the insurmountable challenge in such an endeavor and clearly identified the problems that continue to plague attempts to assess excellence in teaching:

> Unfortunately, it is not as easy to validate the evaluation process as it is to validate a college aptitude test due to the absence of a generally agreed upon criterion or definition of good teaching. In my opinion, it will be impossible to find a generally agreed upon criterion of good teaching. There are probably as many definitions of good teaching as there are teachers. Consequently, in light of the reliability evidence, it appears that we are measuring ‘something’ about faculty members and the way they conduct their classes; however, in the absence of a criterion, it isn’t clear if that ‘something’

> is good teaching or what it is. It should be emphasized that this difficulty is not unique to the present evaluation process. In the absence of a generally agreed upon criterion of good teaching, it will be impossible to directly validate even the most sophisticated evaluation process.
> (Colliver, 1972, pp. 18-19)

Despite the inevitable vagaries of construct definition, Colliver did evaluate the two items using two external criteria: a five-item perceptions-of-teaching form developed by the Berkeley Center for Research and Development in Higher Education (Hildebrand, Wilson, & Dienst, 1971) and ratings provided by SSU Retention Committee members based on prior review of evidence for teaching quality found in faculty personnel files in 1971. The Berkeley form was developed by collecting both student and faculty peer ratings of the "best" and "worst" teachers over a five-year period.

Colliver found that the SSU/UIS two-item mean and the Berkeley five-item mean were positively and significantly correlated (correlation coefficients were .66 and .77; $p < .01$ for each). Thus, he demonstrated content validity for the SSU/UIS two-item rating form. Unfortunately, comparisons with the SSU Retention Committee's qualitative ratings could not be conducted because the Committee members' ratings were found to be too unreliable to use. In other words, a preliminary statistical evaluation of consistency across raters indicated no agreement – the qualitative ratings were unreliable and therefore unusable as an external criterion. According to Colliver:

> Many people have argued for written qualitative student evaluations of faculty which would be read and interpreted by a committee. The present data illustrate what is commonly found to be a problem with such a procedure: the lack of agreement or reliability of the interpretations of the qualitative statements. If student data are to be used to make quantitative decisions (salaries are quantitative), it seems more reasonable to do the quantification at the student level rather than have an intermediate committee make quantitative decisions on qualitative input from students. The latter procedure allows for the subjective biases of the committee members to enter into the decision-making process. (1972, p. 23)

Colliver also sought to assess the validity of several additional faculty concerns about the potential for non-teaching issues to impact student perceptions of teaching quality. He specifically explored the possibility that student ratings might be related to size of course enrollment, course format (lecture vs discussion), teacher use of innovative techniques, academic rank of professor, professors' salaries, number of times course was taught, and highest level of education (e.g., MA/MS, ABD, PhD). The results of correlational analyses showed no relationship between these factors and student ratings. Colliver also examined the relationships between student factors, such as expected grades, age, sex, and undergraduate vs graduate student status and student ratings of teaching. Again, there were no statistically significant relationships – a finding that indicates student perceptions of teaching were differentially unrelated to these student factors.

Current Psychometric Characteristics of the UIS Rating Form Items

At some point in the history of the UIS Rating Form, the wording of Items 8 and 10 were modified slightly and now read as follows (modifications from the original are in bold):

#8. *Do you think this teacher is competent in the content or **material** offered in this course?*

Exceptionally competent		**S**atisfactory		**I**ncompetent
5	4	3	2	1

#10. *Overall, **how** do you **rate the quality of this person as a** teacher?*

excellent	**very good**	good	**fair**	poor
5	4	3	2	1

The items and rating scales remained essentially the same. Adding scale point descriptors for ratings 4 and 2 is consistent with contemporary recommendations regarding the labeling of response point indicators (Arreola, 2000).

Given the length of time that has elapsed since the items were originally developed, the Summative Teaching Evaluation Committee compiled descriptive statistics and conducted a series of analyses using recent data. In order to assess relative stability of aggregate data over time, two subsets of recent data were analyzed separately. One group of ratings for Items 8 and 10 represented the three-year interval of 1995 to 1997, and a second group represented the three-year interval of 2003 to 2005[2]. For all year groups and both items, student ratings are heavily skewed toward the higher end of the rating scale, which means that students tend to rate their course faculty positively or more favorably. Such distributions are technically called *negatively skewed distributions* because there are relatively few ratings at the low end of the scale. Accumulated research, conducted using different rating systems, has shown that student ratings of teaching performance are typically skewed in this way (Arreola, 2000; Cashin, 1999)[3].

Figure 1 presents the 2003-05 group frequencies of average ratings for faculty on Item 8 (the item on which students indicate their perceptions of teacher competence). Only 16% of faculty received ratings at or below 4.0 on the scale. Sixty-eight percent of faculty received average ratings between 4.01 and 4.76, and another 16% had average competence ratings over 4.76.

***Figure 1*: Teacher Competence Rating Frequencies for the 2003-05 Year Group**

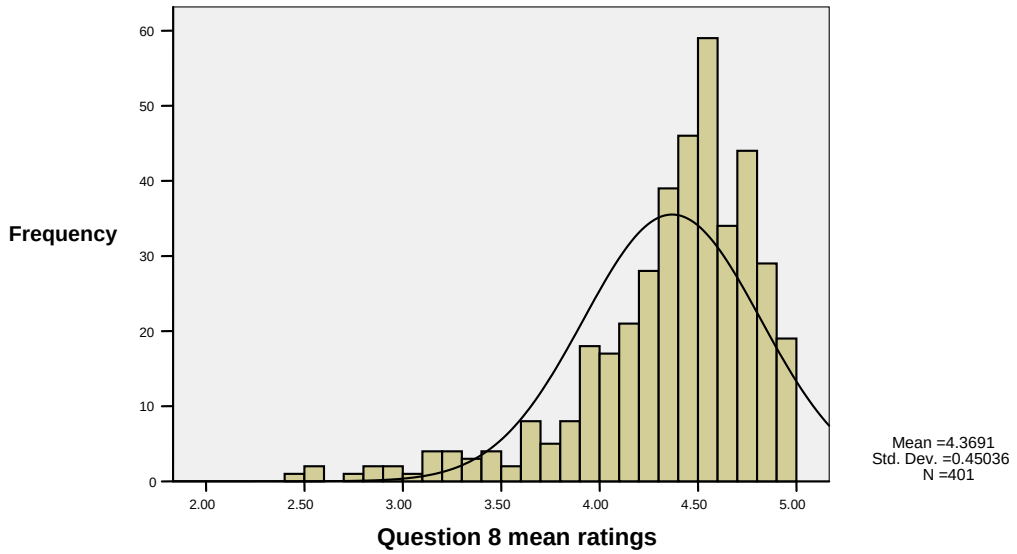


Figure 2 presents the 2003 to 2005 group frequencies for Item 10, which represents average student perceptions of overall teaching quality. Only 16% of faculty received ratings of 3.73 or less. Sixty-eight percent received ratings between 3.73 and 4.70, and the remaining 16% received ratings above 4.70.

***Figure 2*: Overall Quality Rating Frequencies for the 2003-05 Year Group**

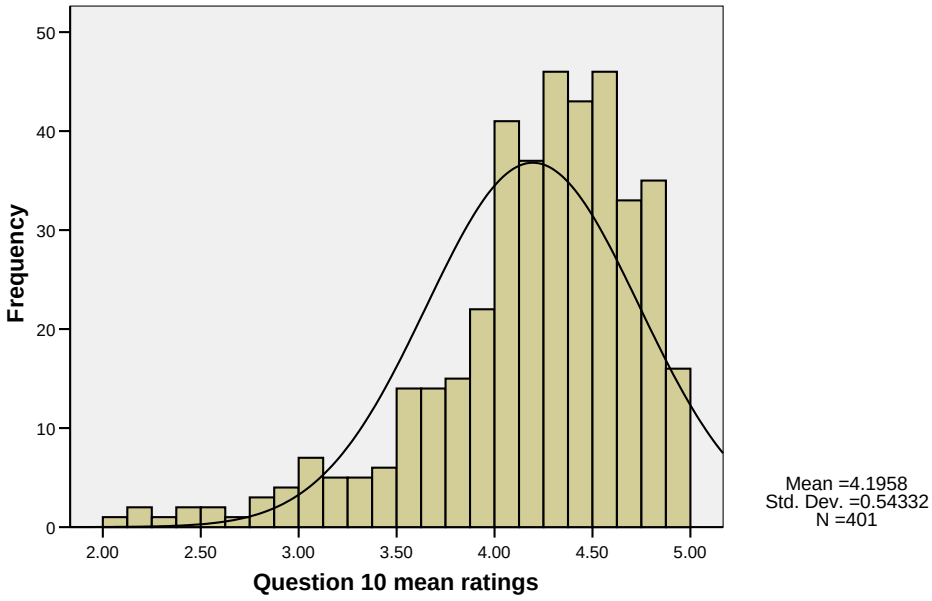


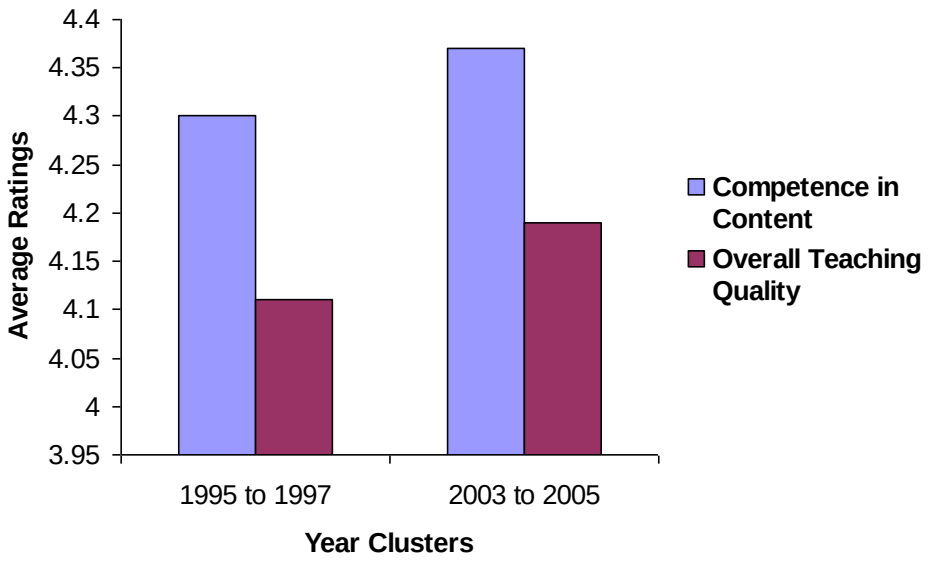
Table 1 presents the descriptive statistics for the recent year groups, as well as for the original 1971 to 1972 and 1972 to 1973 years.

Table 1: Comparative Descriptive Statistics

	Content Competence (Item #8)			Teaching Quality (Item #10)		
Year Groups	***Mean***	***Std Dev***	***N***	***Mean***	***Std Dev***	***N***
2003-05	4.37	0.45	401	4.19	0.54	401
1995-97	4.30	0.45	397	4.11	0.55	397
1972-73	4.51	0.30	179	4.09	0.44	179
1971-72	4.63	0.28	112	4.40	0.37	112

Although there are relative differences between recent rating means and those found in the original years of development, the differences are minor and well within expected shifts over time, given institutional and population changes. There is a statistically significant difference between aggregate ratings across the most recent year groupings, with the average ratings of teacher competence ($t(796) = -2.25$, $p < .05$) and of overall teaching quality ($t(796) = -2.06$, $p < .05$) being slightly higher for the 2003-05 year cluster than for the 1995-97 year cluster (see Figure 3). However, it is important to note that the actual point difference is relatively small (.07 for competence and .08 for quality; Cohen's $d = .15$ for each[4]) and does not necessarily have practical or applied significance. It is not uncommon to detect small but less-than-useful differences when analyzing very large data sets[5].

***Figure 3*: Student Perceptions of Teaching for Recent Year Clusters**



With regard to the reliability of the items, the standard errors of the mean were as follows:

Table 2: Comparisons of Error Within Year Groups

	Standard Error of the Mean	
Year Groups	**Competence** (Item #8)	**Quality** (Item #10)
2003-05	.02	.03
1995-97	.02	.03

Note: *n* = 401 for 2003-05, and *n* = 397 for 1995-97.

It can be seen that random error is very low for both Items 8 & 10, and is stable over the year groups. Thus, the hypothetical 'true' grand mean on Item 8, for any faculty member, will be included within an interval defined by the faculty member's obtained grand mean plus or minus .02 about 68% of the time. The hypothetical 'true' grand mean on Item 10, for any faculty member, will be included within an interval defined by the faculty member's obtained grand mean plus or minus .03 about 68% of the time. Increasing the confidence interval to 96% would increase the standard error of measurement to only plus or minus .04 for Item 8 and .06 for Item 10. It will be noted that the error is lower in more recent samples than in Colliver's original calculations; it is likely that the error is reduced because a much larger sample of course data is represented by the recent three-years clusters. Ceteris paribus, the larger the sample of items included in a score, the higher the reliability of the measurement.

A second technique was used to estimate item reliability for comparison purposes. A series of three consecutive course means, for each faculty member, were entered as items in an analysis of internal consistency. Essentially, in each analysis three consecutive course means, for each faculty member, were treated as separate items on a single scale. Reliability estimates, calculated using Cronbach's alpha, ranged from .87 to .96 for Item 8, and from .80 to .85 for Item 10. Such coefficients are very high, and indicate that aggregate data for these two items demonstrate remarkable consistency across courses for any given faculty member.

Intercorrelations of recent ratings for competence and overall quality were calculated within years, and, consistent with Colliver's original report, the present results were statistically significant and quite strong

(Pearson *r* ranged from .86 to .91, with *n* ranging from 226 to 273; see Table 3). As perceived competence increases, ratings of overall quality as a teacher increase.

Table 3: Intercorrelations Between Items 8 & 10 Within Years

	Quality						
Competence							
	1995	***1996***	***1997***		***2003***	***2004***	***2005***
1995	.89						
1996		.86					
1997			.86				
2003					.90		
2004						.91	
2005							[a]

Note: all coefficients significant at the .01 level; *n* ranged from 87 to 194. [a] Only Spring semester data were available, so comparable coefficients could not be calculated.

Intercorrelations for competence and quality *across* courses and semesters were predictably somewhat lower[6], but were statistically significant and indicated moderate levels of association even over time (see Tables 4 & 5). Thus, earlier ratings of competence and teacher quality continue to be somewhat predictive of later ratings of competence and teacher quality over time.

Table 4: Ratings of Content Competence: Average Correlations Across Years

Years					
	1996	***1997***		***2003***	***2004***
1995	.51	.49		.35	.38
1996		.51		.37	.21
1997				.32	.39
2003					.45

Note: all coefficients significant at the .01 level; *n* ranged from 87 to 194.

Table 5: Ratings of Teaching Quality: Average Correlations Across Years

Years					
	1996	***1997***		***2003***	***2004***
1995	.48	.47		.30	.52
1996		.57		.36	.32
1997					.45
2003					.44

Note: all coefficients significant at the .01 level; *n* ranged from 87 to 194.

Student Perceptions of Instructor's Presentation & Organization

Item #7 on the current student ratings form asks students to indicate whether they think their instructor's presentation was well-planned and organized or not (yes/no response option). Overall, the vast majority of students, in both the 1995-1997 and 2003-2005 data subsets, indicated their instructors' presentation was well-planned and organized (91% for both subsets). Not surprisingly, students who indicated their instructor's presentation was well-planned also rated their instructors' content competence and overall teaching quality significantly higher than did those students who did not view the presentation as well-planned. For the 2003-05 data subset, perceptions of instructor's presentation predicted 26% of the variability in student ratings of content competence and 40% of the variability in student ratings of teaching quality (assessed using Phi coefficients).

Table 6: Average Ratings by Presentation Category

Presentation	Teacher's Content Competence		Teaching Quality	
	1995-1997	2003-2005	1995-1997	2003-2005
Well-Planned	4.5	4.5	4.4	4.4
Not Well-Planned	3.2	3.1	2.6	2.5

Note: all differences between Well-Planned and Not Well-Planned were evaluated using an independent samples *t*-test and are significant at the .001 level. For the 2003-2005 year group differences, Cohen's *d* is 1.88 for competence and 2.42 for quality[3].

Student Motivation and Critical Thinking

Items #6 and #9 on the student ratings form asks students to indicate how the course affected their performance motivation and critical thinking skills (yes/no response options). A majority of students, in both the 1995-1997 and 2003-2005 data subsets, indicated their courses had motivated them to work at their highest levels (78% for both subsets) and increased their critical thinking skills (86% and 83%, respectively). Students who indicated the course had motivated them and increased their critical thinking skills also rated their instructors' content competence and overall teaching quality significantly higher than did students who indicated the course had not motivated them or increased their critical thinking. For the 2003-05 data subset, motivation predicted 19% of the variability in student ratings of content competence and 28% of the variability in student ratings of teaching quality (assessed using Phi coefficients). Critical thinking predicted 15% of the variability in student ratings of content competence and 23% of the variability in student ratings of teaching quality.

Table 7: Average Ratings by Effect Category

Self-Reported Effect	Teacher's Content Competence		Teaching Quality	
	1995-1997	2003-2005	1995-1997	2003-2005
Motivation				
Motivated	4.6	4.6	4.5	4.5
Not Motivated	3.8	3.7	3.3	3.3
Critical Thinking				
Increased	4.5	4.6	4.4	4.5
Not increased	3.7	3.6	3.1	3.2

Note: Differences within Self-Reported Effect categories are significant at the .001 level (*t*-test). For the 2003-2005 year group differences, Cohen's *d* ranged from 1.18 to 1.53[4].

Analyses to Assess for Associations with Extraneous Variables

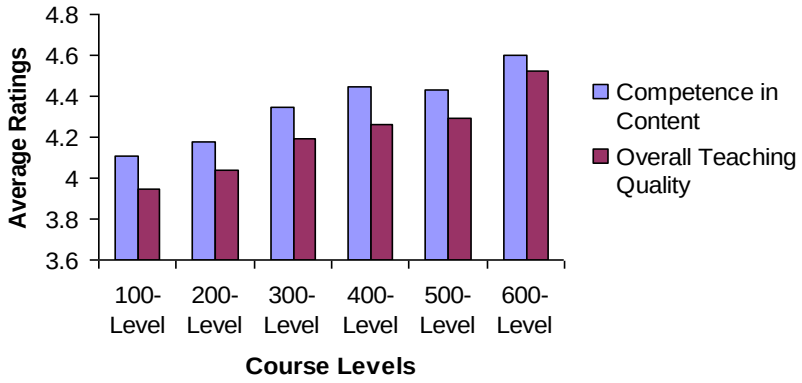
Separate statistical analyses were conducted to assess for the possibility that faculty gender, student gender, course level, required vs. elective enrollment, and expected grade are associated with student ratings within the recent three-year cluster. Tests to determine whether there might be significant associations with part-time vs full-time faculty status, tenure status, or actual grade received could not be conducted because, historically, these variables have not been included in the database, and could not be easily added solely for the purpose of this report. Although course prefix and number are recorded in the data base, no analysis for possible academic area effects was possible because program area couldn't be identified for several lower-division courses that carried only the Capital Scholars Program prefix.

Faculty Gender
Faculty gender has not been recorded in the archived data set and had to be added to the data subsets selected for the present analyses. No statistically significant differences in student ratings were found for faculty gender for either the 1995-07 or the 2003-05 data subsets.

Student Gender
T-tests indicated that quality ratings given by women students were slightly lower than those given by men students within the 1995-97 group (4.23 vs. 4.26, respectively; (t(41933) = -.03, $p < .01$). For the 2003-05 group, women students gave slightly higher competence ratings (4.43 vs. 4.40 from men; (t(31806) = 2.76, $p < .01$; Cohen's d = .03) and higher quality ratings (4.27 vs. 4.24 from men; (t(31819) = 2.04, $p < .05$; Cohen's d = .03). Even so, it should be noted that, once again, the sheer number of cases in the data set allowed for an analysis statistically powerful enough to detect even small differences in ratings. Student gender accounted for less than 1% of the variability in student ratings for the 2003-05 data subset.

Course Level
A one-way analyses of variance (ANOVA) was conducted on the 2003-05 group only because it included a full range of undergraduate and graduate course levels. Results indicated that there were significant differences in student ratings across course levels for both content competence (F (5, 786) = 7.03, $p < .001$) and overall teaching quality (F (5, 786) = 4.53, $p < .001$). As shown in Figure 4, student ratings tend to increase as course level increases. This finding is consistent with the broader literature; according to Arreola, "there appears to be a tendency for graduate or upper division students to rate instructors more favorably than lower division students" (2006, slide 80). In practical terms with regard to item strength, course level accounted for only 2% of the variability in student ratings (r (1506) = .15, p < .01 for competence; r (1506) = .14, p < .01 for overall quality).

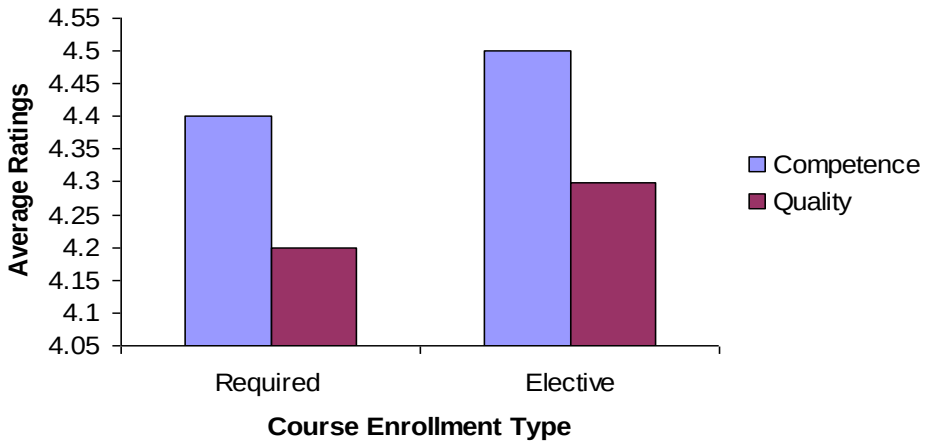
***Figure 4*: 2003-05 Student Perceptions of Teaching by Course Level**



Required versus Elective Enrollment

Item #4 on the current student ratings form asks students to indicate whether they enrolled in the course as an elective or because it was required. Approximately 70% of students indicated they were taking the course because it was required. A t-test indicated students who reported the course was an elective gave slightly higher competence ratings than students who reported the course was required (4.48 vs. 4.39, respectively; $t(31993) = 9.17$, $p < .001$; Cohen's $d = .10$). Students who reported the course was an elective also gave slightly higher teaching quality ratings than students who reported the course was required (4.33 vs. 4.22, respectively; $t(32006) = 8.96$, $p < .001$; Cohen's $d = .11$). However, course enrollment type accounted for less than 1% of the variability in student ratings of competence and quality.

***Figure 5*: 2003-05 Student Perceptions by Course Enrollment Type**



This finding is consistent with the accumulated literature on student ratings, which indicates that students tend to give higher ratings when the course is taken as an elective.

Expected Grade

Expected grade codes (Item #3) were converted to grade point format for analyses. There were no statistically significant differences in the average expected grades for the 1995-97 and 2003-05 data subsets. Average expected grades were subsequently analyzed relative to student ratings of teacher competence and quality within the 1995-1997 and 2003-05 data subsets using Pearson correlation. Higher expected grades were significantly associated with higher ratings of competence and quality for both year clusters. For competence, *r*s were .12 and .14, $p < .001$, respectively. For overall quality, *r*s were .19 and .20, $p < .001$, respectively. Nevertheless, the actual impact on ratings was relatively small. Average expected grade was associated with less than 2% of the overall variability in competence ratings, and with approximately 4% of the overall variability in quality ratings. In other words, 98% of the variability in ratings of competence and 96% of the variability in ratings of quality were *not* associated with expected grades.

It is important to emphasize the distinction between ***expected*** grade and ***given*** grade because one of the strongest and most persistently cited concerns about the use of student ratings of teaching is that their use in the personnel process routinely leads to spuriously high grades – presumably resulting from instructors trying to please their "reviewers" in order to receive high ratings. The possibility that student perceptions of teaching could be unduly influenced by evaluations of their own course-related performance largely dates back to experimental research conducted in the 1970s, in which grading in a sample of undergraduate courses was directly manipulated to determine the effects on student ratings of teaching (see Greenwald, 1997). The results of that early research seemed to indicate that student ratings could be related to obtained grades; however, the studies themselves were later criticized heavily for methodological problems. In the ensuing two decades, considerable research attention was given to investigating the potential relationship

between grading and ratings. As investigations and analyses have become increasingly sophisticated, it has become clear that student ratings of course and teacher quality represent a complex amalgam of many variables, and are associated with measurable student learning, teaching-related behaviors of the faculty, students' pre-course interest in the subject matter, and course level (Marsh & Roche, 1999; Schneider, Hanges, Goldstein, & Braverman, 1994). Expert opinion on the subject has largely shifted, and a significant number of leading researchers in the field today do **not** believe that there is a substantial or consistent relationship between assigned grades and ratings that reflect student perceptions of teaching. For example, consider the following:

> …we know of no persuasive evidence that establishes that instructors who assign unjustifiably high grades receive glowing ratings of instruction as a result. Our own meta-analytic review of field experiments…suggests that the results are small on average, variable, and not readily separable from the *valid* influences of instructors on student learning. In one experiment, poor teaching coupled with higher grades resulted in more negative ratings. (Abrami & d'Apollonia,1999, pp. 519-520)

> Grading leniency is not easily operationalized because class-average expected grades do not represent grading leniency unless preexisting student-course characteristics and achievement are controlled (Classes with better students who learn more should receive higher grades). The SET [student evaluations of teaching]-expected grade relation is small (r = .2 for overall teacher ratings), and there are many empirically supported explanations of this relation that do not implicate biases. Contrary to bias hypotheses (a) overall course ratings are more correlated with grades than overall teacher ratings, (b) some SET factors (Organization, Enthusiasm, Breadth) are unrelated to grades, (c) controlling prior characteristics (course level, prior subject interest) and achievement substantially reduces the relation, and (d) in support of SET validity, the highest correlation (.3) is with learning-value ratings. Furthermore, the SET-grade relation is nonlinear (nearly flat above the mean grade with a small decline for extremely high grades) so that high grades are unrelated to SETs….Although grading leniency might explain a small portion of a small relation, there is little empirical support for this speculation. (Marsh & Roche, 1999, 517-518)

Summary and Recommendations

Instructional evaluation, as a central focus in the evaluation of faculty performance, is a complex and high-stakes topic. Best practices in higher education consistently point to the need for gathering multiple perspectives and sources of information when making evaluative judgments that will contribute to personnel decisions. Clearly, student perceptions of teaching constitute only one source of information; however, national surveys also indicate that student ratings have become the most commonly used source of information about teaching. A 1998 survey of all accredited, four-year, liberal arts colleges in the United States asked College Deans to indicate which of 15 different sources of information were used in assessing the teaching effectiveness of faculty at their institutions (Seldin,1999). The survey response rate was strong, with 81% of the colleges responding. Four sources of information were cited as being consistently used by more than 50% of the responding institutions: systematically gathered student ratings (88.1%), evaluation by the department chairperson (70.4%), evaluation by the dean (64.9%), and self-evaluation (58.7%). Approximately 76% of the responding institutions indicated they use a rating form to gather student, colleague, and/or administrator feedback; however, only 14.4% reported that the institution had conducted research to assess the validity of the forms they use.

Although student ratings are the most common component of faculty evaluation approaches, as previously noted, experts continue to caution against over-emphasizing the value of student ratings and recommend that multiple sources of data be used to assess teaching quality (e.g., Arreola, 2000). For example, Cashin (1999) explicitly points out that teaching involves more than what students observe or experience in the classroom, and students are not able to objectively evaluate an instructor's level of expertise, rationale for the curricular plan and course design decisions, or the complexities of classroom management. Even so, having acknowledged the realistic limits of student feedback, there is still a consensus that well-constructed rating forms can be an important and useful component of the faculty evaluation process. Students can provide valuable information about how they have experienced the learning process itself in terms of the

pedagogical methods and modes of instruction, accessibility and learning support, assessment exercises, and course administrative structures. The key to gathering useful student feedback is to be certain that information is collected consistently using a well-developed, reliable, and valid measure.

The UIS Student Rating Form was originally constructed to ensure the reliability and validity of the critical summary items. Comparative analysis of the original descriptive statistics and those for more recent years suggest strong stability of item functioning. Although there are relative differences between recent rating means and those found in the original years of development, the differences are minor and well within expected shifts over time, given institutional and population changes. More importantly, analyses of aggregated contemporary data sets clearly demonstrate the reliability and resilience of the two items that assess student perceptions of competence and quality. These are important foundational considerations for making evaluative comparisons across faculty, and our findings support the continued use of the current UIS rating form as one important summative component in the evaluation of teaching.

While the accumulated research on well-developed ratings forms indicates that they can provide an important source of information about the quality of teaching, a number of common concerns about the validity of student rating forms continue to be raised. Such concerns typically take the form of assertions such as:

(1) "Student ratings just reflect how popular the teacher is."
(2) "Students will give higher ratings to teachers who give easy As."
(3) "Student rating forms are gender-biased; students give women faculty lower ratings than men
(4) faculty."
(5) "Student ratings are affected by factors like class size."
(6) "Students aren't capable of making informed judgments about my teaching."

Aleamoni (1999) now refers to these concerns as "myths" because they are not supported by the accumulated literature on reliable and valid rating forms (see Arreola, 2000, and Abrami & d'Apollonia, 1999, for succinct summaries of the accumulated literature). Well-developed rating forms -- whether formative or summative -- should be reliable and valid measures of several aspects of teaching, including general attitude, organizational effectiveness, engagement, and teaching skills and characteristics. Student ratings are not typically associated with factors like class size, the gender of the instructor, or grades received. The literature does, however, identify three factors that have been found to be consistently associated with student ratings: course level, elective enrollment status, and academic discipline. Students in upper-division undergraduate courses tend to rate their professors more favorably than students in lower-division undergraduate courses do, and students in graduate-level courses tend to rate their professors more favorably than students in undergraduate courses do. Students also tend to give higher ratings when they have taken the course as an elective, rather than a requirement. And there is some evidence that student ratings differ by academic discipline, with ratings for courses in the humanities and social sciences being more favorable than those in math and physical sciences.

Analyses of recent data for the UIS Student Rating form are quite consistent with the broader literature. Student ratings at UIS were not associated with faculty gender, and only inconsistently associated with student gender. UIS student ratings were associated with course level, enrollment status, and expected course grade; however, associations were quite small in magnitude and of no practical value. For the 2003-2005 subset, course level predicted approximately 2% of the variability in student perceptions of content competence and overall teaching quality, enrollment type predicted less than 1% of the variability in perceptions of competence and overall teaching quality, and expected grade predicted 4% of the variability in content competence and 7% of the variability in teaching quality.

In order to assess the relative predictive contributions of course-related variables (such as presentation planning and effects on motivation and critical thinking skills) versus extraneous variables (such as course enrollment type, course level, and expected grade), associated variables were entered into a multiple regression analysis. Results indicated that the strongest predictors of both perceived content competence and teaching quality for the UIS student ratings form were, in order of predictive magnitude, student perceptions of instructor presentation, course impact on student motivation, and course impact on student critical thinking. This combination of variables predicted 31% of the variability in student ratings of

instructors' content competence (Item #8) and 46% of the variability in student ratings of teaching quality (Item #10). Thus, student ratings of teaching using the current UIS form are more strongly related to perceptions of course-related characteristics than to variables unrelated to the course itself (i.e., than to extraneous variables that are typically cited by faculty as potential sources of student bias in ratings).

Recommendations for Data Set Management and Interpretation

Overall, the results of our analyses indicate that the current UIS student rating form demonstrates acceptable levels of reliability and validity for continued summative use as one measure of teaching quality. In addition to research on the development of student rating forms, Arreola (2000) provided a set of "best practices" guidelines for administering the form and processing the data collected. Along these lines, it should also be noted that the procedures at UIS are consistent with these recommendations with regard to administration of the student rating form and with regard to taking a multifaceted approach to the evaluation of teaching. In addition to having a clearly specified process by which student evaluation forms are administered, collected, tabulated, and accessed, UIS has a faculty personnel policy that explicitly makes data derived from student ratings of perceived teaching quality only one source of information about teaching quality. This, too, is consistent with expert recommendations. For example, Cashin (1999) writes:

> In the main, past and present reviewers continue to conclude that student ratings tend to be statistically reliable, valid for most uses, relatively free from bias or the need for control, and useful both to improve instruction and to make personnel decisions. However, regarding personnel decisions, there is almost universal agreement that data from a variety of sources, not just student ratings, are required to accurately evaluate teaching. (p. 28)

Even so, as Cashin points also out, a psychometrically sound, summative rating form can be a useful component of the personnel review process when used with interpretive reference to appropriate normative data:

> The average student rating on a five-point scale is not 3.0—as one might think—but usually between 3.5 and 4.0. Also average ratings vary widely from item to item. On the 20 IDEA teaching method items, the lowest mean is 3.4; the highest, 4.4. What should an instructor conclude who receives an average rating of 3.9 on these two items? Without comparative data, I do not believe that it is possible to meaningfully interpret student rating data. (p. 32)

While direct comparisons across different student rating forms cannot be made, aggregated UIS averages on Items #8 and #10 of our student ratings form indicate that the vast majority of students in UIS courses regard their faculty as excellent teachers.

Nevertheless, there are several ways in which ***the use of student ratings*** could be improved at UIS, and we offer the following recommendations toward that goal:

1. ***Develop standard basic normative reports using appropriate subsets of the total data base for use in the personnel process.***

   Although UIS has accumulated an extensive data set of student ratings, it is recommended that only recent and relevant years' data be used in the personnel process. It is important to note that, although the aggregate means demonstrate stability over recent years, the relationship between percentiles and average ratings inevitably shifts slightly across year groupings. For example, in the 1995 to 1997 year cluster, 68% of faculty received average student ratings between 3.93 and 4.73 on teacher competence, while in the 2003 to 2006 year cluster, the same percentage of faculty received average student ratings between 4.01 and 4.77. While such changes are relatively subtle, they could, conceivably, have relevance for the interpretation of student ratings inasmuch as they indicate the potential importance of compiling normative data across an appropriate range of years. Ideally, normative data should be compiled for the cluster of semesters against which any given faculty member's average has been calculated.

Aggregated data from the previous academic year and from the most recent six-year period should be used to develop two annual interpretive reference reports. Each report should include, in the most user-friendly format possible, percentile scores, grand means, standard deviations, standard error of the grand means, and modes for Items #8 and #10. Data for individual faculty should be aggregated before calculating the statistics for the overall group, and all average scores should be presented with relevant confidence intervals in recognition of expected ranges of fluctuation (i.e., standard error).

The six-year report would, in most instances, constitute the best normative reference for interpreting student ratings as part of the typical tenure-review process (assuming no eligible teaching experience prior to UIS) because it captures more information than is available at the point of two- or four-year reviews. The reliability of mean ratings increases as the number of semesters of ratings increases; therefore, average ratings of teaching competence and overall quality are less reliable and predictive at the point of two-year review than at the point of four-year or six-year review. Stated somewhat differently, ceteris paribus, the greater the number of student ratings, courses, and semesters represented by the mean, the greater the confidence one can have in the stability and representativeness of the average rating.

2. ***Develop a set of written interpretive guidelines to accompany the annual normative reports, to help guide all involved in the personnel evaluation process in interpreting student ratings appropriately, with due regard for the nature of the underlying distribution and for standard errors of measurement.*** Comparison data should also be presented in graphed or figural forms as well. For example, frequency distributions of ratings on Items #8 and #10 should be presented in graphed form to facilitate the understanding of percentile scores. Mean scores should be presented in a figural context that includes relative ranges of expected variability (i.e., confidence interval bands based on standard errors of measurement).

3. ***Prepare and conduct annual workshops, voluntary for faculty but mandatory for all personnel committee members and academic administrators, on best practices in the use of student ratings of teaching as one source of evidence in the evaluation of teaching.***

4. ***Develop the current data base to facilitate periodic reviews of the psychometric quality of the student ratings form and to allow for refinement of the form and of the interpretive guidelines over time.*** At a minimum, the following variables should be added to the data set: faculty gender; academic field or cluster (i.e., humanities, science, social science, etc.); course level indicator; full-time/part-time faculty status; and tenure status. Such variables will allow for periodic monitoring of differences that might be related to non-teaching variables (such as student gender or course level), as well as for the selection of specialized comparative groups if needed. For example, in evaluating student perceptions of teaching for a faculty member whose course load has consisted largely of lower-division general education courses, one might want to consider percentile equivalents based on the aggregated evaluations for all lower-division general education courses taught at UIS during the relevant review period.

Bibliography

Abrami, P.C., and d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist, 54*(7), 519-520.

Aleamoni, L.M. (1999). Student rating myths versus research facts: An update. *Journal of Personnel Evaluation in Education, 13*(2), 219-231.

Arreola, R. A. (2006). *Student ratings: Their design, construction, and use*. Paper presented at the audio conference on March 9, 2006. http://www.magnapubs.com/calendar/index-cat-type.html#audioconf Madison, WI: Magna Publication, Inc.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker Publishing, Inc.

Basow, S.A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656-665.

Cashin, W.E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin & Associates, *Changing practices in evaluating teaching (pp. 25-44)*. Bolton, MA: Anker Publishing, Inc.

d'Apollonia, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208.

Freeman, H.R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology, 86*(4), 627-630.

Gillmore, G.M., & Greenwald, A.G. (1999). Using statistical adjustment to reduce biases in student ratings. *American Psychologist, 54*(7), 518-519.

Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186.

Greenwald, A.G., & Gillmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology, 80*(3), 342-344.

Marsh, H.W., & Roche, L.A. (1999). Rely upon SET research. *American Psychologist, 54*(7), 517-518.

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

McKeachie, W.J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*(11), 1218-1225.

Schneider, B., Hanges, P.J., Goldstein, H.W., & Braverman, E.P. (1994). Do customer service perceptions generalize? The case of student and chair ratings of faculty effectiveness. *Journal of Applied Psychology, 79*(5), 685-690.

Seldin, P. (1999). Current practices—good and bad—nationally. In P. Seldin & Associates, *Changing practices in evaluating teaching (pp. 1-24)*. Bolton, MA: Anker Publishing, Inc.

Thalheimer, W., & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved January 17, 2007 from http://work-learning.com/effect_sizes.htm.

Notes

[1] Faculty names were removed from the data subset before it was shared with the Summative Teaching Evaluation Committee; thus, all cases analyzed for the purposes of this report were identified by a code number only.

[2] The 1995 to 1997 years were chosen, in part, because they represented the first years under the University of Illinois at Springfield designation. With regard to the more recent 2003 to 2005 year group, data analysis for this report began in the spring of 2006, so 2005 represents only the available fall semester data.

[3] Several faculty members have asked Dr. Hadley-Ives whether a "good" measure of student perceptions of teaching should yield a "normal" or bell-shaped distribution of ratings. By implication, the questioners are suggesting that the current measure of student perceptions cannot be psychometrically sound if it yields ratings that are negatively skewed. Readers should note that there are many psychological dimensions and behavioral characteristics that, when assessed quantitatively, normally yield *non-normal* (i.e., skewed) distributions; thus, the fact that a scale yields a non-normal distribution does not necessarily reflect a psychometric flaw in the scale's development. Dr. Hadley-Ives has written a special short paper to try to explain, as simply as possible, why the current rating scale is appropriate and useful despite the fact that student ratings of teaching naturally fall into a negatively skewed distribution. His paper also touches on the subject of construct complexity and validity, and is included with this report, as Appendix A.

[4] Cohen's *d* is a measure of effect size and represents group differences in average scores as a proportion of pooled variance. Recommended interpretative guidelines categorize effect sizes of approximately .20 as small, .50 as medium, and .80 as large (see Thalheimer & Cook, 2002, for an overview).

[5] Just as microscopes that have greater powers of magnification are able to detect a smaller level of cellular detail, statistical tests conducted on larger data sets have greater power to detect small group-level differences in scores. The meaningfulness of the differences detected depends on relative effect size and other practical indicators.

[6] For any given faculty member, student ratings of teacher competence and quality can be expected to vary across courses and time. Associations will necessarily be weaker as the number of course repetitions decreases and as the time that elapses between the ratings increases. Nevertheless, Tables 4 and 5 indicate that earlier ratings of competence and teacher quality continue to be predictive of later ratings of competence and teacher quality.

Appendix A

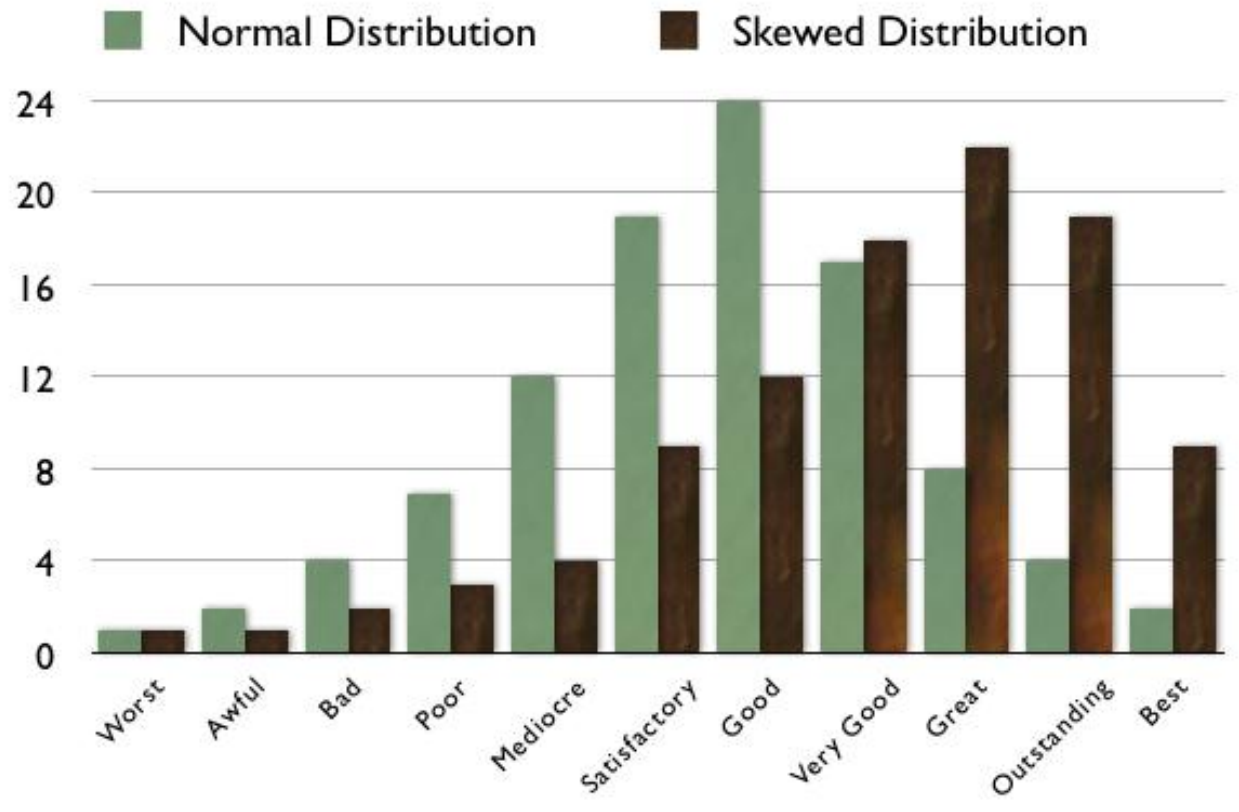
Explanatory Paper by Eric Hadley-Ives

Why a seemingly crude measure is better than a refined measure.

With a range of only 1-5 and anchor points of poor, fair, good, very good, and excellent, it *seems* obvious that item 10 (and the related item 8) are far too crude to give us much useful information about student perceptions of our teaching and competence. Yet, when I reflect upon what a "better" measure would mean, I come back to our "crude" measure and find it isn't so bad after all. I'd rather use the crude measure in comparisons among teachers. A more refined measurement might be helpful to me in improving my teaching, but I wouldn't want to use a "better" scale in comparisons with other faculty because a "better" scale would probably be more sensitive to random noise, and would yield deceptively fine ratings of something that is extremely messy (student perceptions).

Let me explain why this is so.

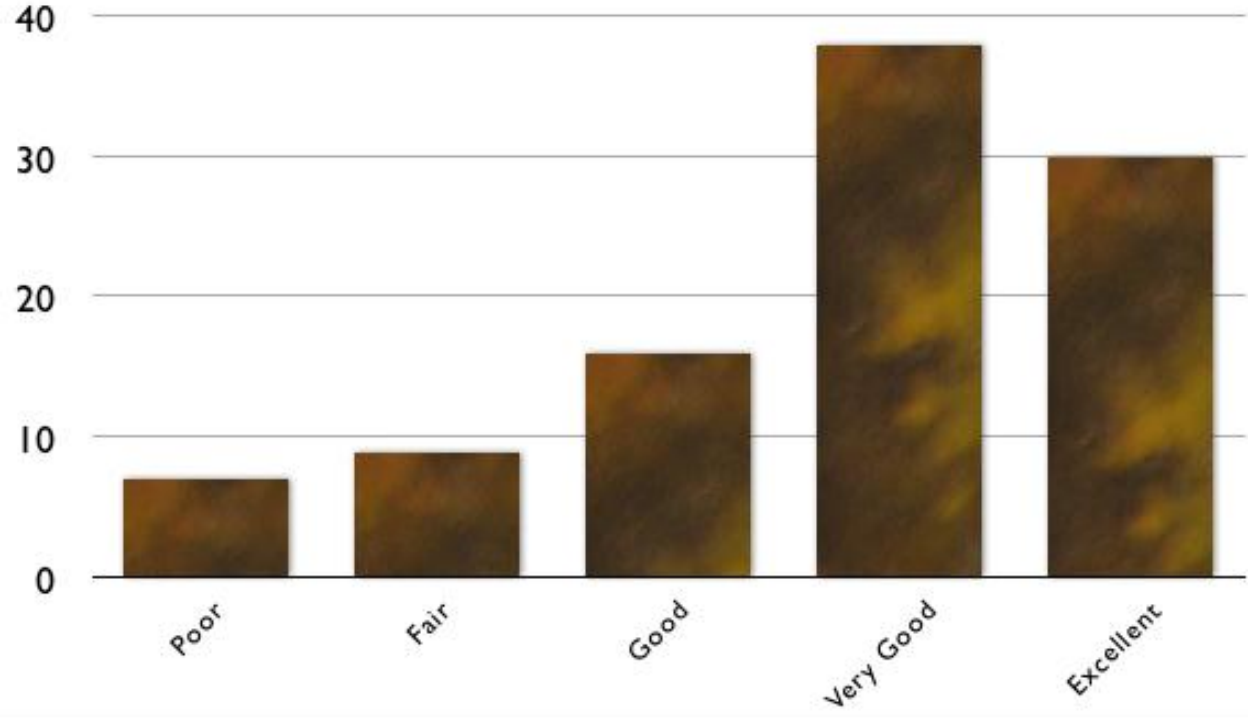
First, let's look at a normal distribution of scores and compare this to the sort of skewed distribution we have in the UIS student evaluations of our teaching. This negative skew is one fact that seems like a problem with what we have with our current scale:



In the above chart I've imagined what sort of evaluations we might get if we expanded the 5-choice scale to an 11-choice scale. The lighter shaded bars represent something like a normal distribution, where most teachers would be rated satisfactory or good or

very good, and few would be rated at the highest and lowest ends of the scale. In fact, when students rate teaching quality they tend to give skewed answers so that more teachers are rated at the higher end of the distribution, as shown in the darker bars in the chart above. Such a skew is found in several sorts of questions. For example, if we ask people about their degree of happiness, or life satisfaction, we find almost exactly the same sort of skew, where the tallest-point of the distribution (mode) is close to the highest possible answer, and the average (mean) rating is somewhere between two-thirds and three-quarters of the way toward the highest end of the possible range. This seems like a problem because there is a "ceiling" at the highest end of the possible responses, and most student ratings are clustered tightly together at the "very good" and "excellent" end of the ratings. If you get "only" a "good" rating it seems that you are scoring too low, and a rating of "satisfactory" is so far below the typical "outstanding" rating that the meaning of "satisfactory" is changed to something like "way too low compared to other ratings and therefore unacceptable."

Rather than using an 11-point rating system, we use items that have only five, three, or two possible ratings. So, student evaluations of our teaching look more like the scores in the following chart:



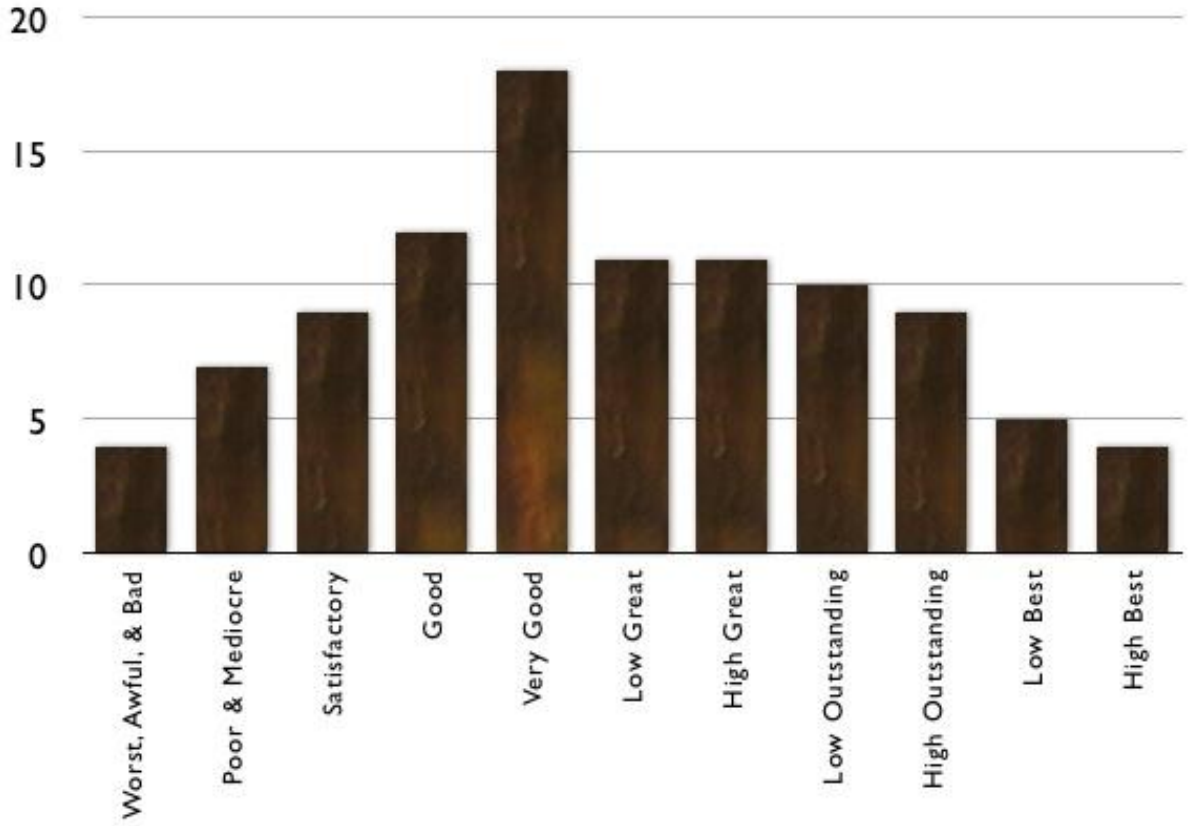
Actually, the distribution in this chart is the same as the darker bars in the first chart, I've just collapsed categories together to reduce the eleven response categories into five.

By having fewer categories it is much more difficult to get statistically significant differences between ratings of "good" and "very good" and "excellent" (3, 4, or 5). This is an advantage when we're comparing our student ratings to some standard. Why is

this? Because the distance between a student's perception of "good" and "very good" may only be a hairbreadths, and might be influenced by what a student had for breakfast or some other trivial or random event. We want to be able to distinguish between poor teaching and good teaching so we can give attention to what is going on in classes where people get poor ratings. We also would like to distinguish between the merely fair ratings and the excellent ratings so those of us who get fair ratings can go observe classes that get excellent ratings and see how to improve. But do we really want to be making decisions about careers and promotion based upon differences among the "good" and "very good" and "excellent" ratings?

Keep in mind that this is an ordinal scale, and there is no way to know the distances between "fair" and "good" or "good and "very good" as there would be if we were measuring something like temperature or physical length with a thermometer or tape measure. Cutting up the categories of "very good" and "excellent" into more categories (as shown in the first chart) might be similar to measuring tiny differences that are more susceptible to trivial or random student perceptions. With more response categories and wider score variations there would be a stronger possibility of finding statistically interesting artifacts. For example, improving the scale might let us discover that two teachers who had previously indistinguishable averages of 4.2 and 4.5 on a scale with a potential range of 1-5 actually have a statistically significant difference of 7.4 and 8.1 when we spread the distribution to a 1-11 range. Such findings might be statistically significant, but they would probably be meaningless in practical terms, and it seems likely that significant differences would arise on a semester-to-semester or year-to-year basis depending upon matters unrelated to our actual teaching quality (such as the particular configuration of personalities in the classroom, or how we were feeling in the last few weeks of the semester leading up to the student evaluations). Scales with greater sensitivity are more likely to find "statistically significant" things, even when there is no practical significance to the findings.

If we wanted to get rid of skew and try to create a measure where a normal distribution would emerge we could change anchor points to stretch out the responses at the good end and collapse answers at the bottom end. The next chart shows the same distribution as shown in the dark bars in chart 1 and chart 2, but this time the response categories are designed to give a more normal distribution:



Such a rating scale would reduce the skew, but would it help us evaluate student perceptions of our teaching? Probably it wouldn't. Such a rating scale might be something like using a yard-stick that has measure markings at 1-ft, 2-ft, and then markings at 2-ft-6-inch, 2-ft-9-inch, 2-ft-10.5 inch, 2-ft-11.25 inch, 2-ft-11.625-inch, 2-ft-11.8125-inch, and so forth to measure the position of a football on a field. The thing that is measured (student perception of our teaching) is too "fuzzy" to be accurately measured by fine distinctions between "low outstanding" and "high outstanding" and so forth, just as we can't position a football at the 10-yard-2-foot-4-and-3-16^{ths} inch line on a football field. All we really need to know is, did we score above the "poor" ratings, and if we're only in the "good" ratings can we move up to something higher than "good"?

A finely graded scale with many divisions between good and "best/perfection" *might* be useful for long-term studies of our own personal teaching. For example, we might try to improve student perceptions of our classes to a point where students go from telling us, "you're a great professor and I really enjoyed your class" to a situation where they say, "you're the best teacher I've ever had and your course has changed my life, I'll never forget this." A finely graded scale with an expanded range of possible positive responses might help us get fairly normal distributions we could examine with statistically sophisticated analysis to document some improvement, but one hardly wants to use such untried and potentially misleading measures to evaluate people for reappointment or tenure decisions, especially if there are only a few semesters of

student responses, and those responses represent only a non-random sampling of students who were most motivated to fill out the evaluation forms.

Until scales with wider ranges and alternative anchor points are carefully tested and developed we're better off sticking with the sort of scales we have been using. We just have to do the following:

- Use the scales to see if a particular person's range of student responses generally overlaps the range of responses other professors at this school get.
- Be sure that when we're comparing a professor's range of student responses we are comparing it to an appropriate university standard. Student ratings from lower division Capital Scholars courses need to be compared to range from other lower division Capital Scholars courses, not overall university averages. Lower-division course ratings should be compared to lower-division averages. Required lower division math and science course ratings should not be compared to averages that are heavily influenced by ratings from upper-division elective courses taken mainly by students majoring in course subject areas.
- We must usually not compare means. It makes no sense at all to say a particular instructor's mean score is "lower" than a department's or school's mean score. The inherent variability in any instructor's scores are unlikely to make differences between means to be interpretable or significant unless the differences are very wide indeed. Any average ratings on our 5-point scales (items 8 and 10) in 4s, or even in the high 3s, almost certainly represent very positive student perceptions. Only when average scores sink to the low 3s or 2s would such averages probably represent a problem, but here again, this would depend upon the type of course. In some types of lower-division courses student ratings in the low 3s might be expected and welcomed. After all, the anchor point for a "3" is "satisfactory" (on item 8) or "good" (on item 10).

I am trying to develop "better" scales for my own formative evaluations of my teaching and student perceptions of my teaching. I'm eager to share these scales with colleagues, and would be especially enthusiastic about combining the sort of student ratings we receive so I can evaluate the psychometric properties of these alternative formative evaluation instruments. If you are interested, please contact me. If you want to see an example of the sort of scale I'm playing with, check out a sample at people.uius.edu/hadleyiv/eval.html.

Anther objection to the evaluations we use is that they rely too much on student perceptions of teaching quality, rather than more objective measures of teaching quality. As our report highlights, we don't have a good criterion to use as a comparison to test the validity of the student perceptions of teaching excellence. What would such a criterion look like?

Probably the only way to generate a good measure of teaching quality is to use a multi-trait-multi-method approach. We would use pre-tests and post-tests of course content

to show knowledge gain. We would track students after they had graduated to see whether their learning in a course had been usefully applied in their careers, and whether they had increased salary or satisfaction or success through talents or knowledge they had picked up in our classes. We would have trained raters observe videos of our classes to count instances of "best practices" in teaching or score all the unhelpful things we do and say with our students. We would have our students take discipline-specific qualifying exams and note any correlations between who had taught them particular courses in their disciplines and how they had performed on particular areas of their qualifying exams. We would interview students to ask them about their experiences in our classrooms. We would control for variables such as students' scores on intelligence tests, the life burdens that were distracting them from classes when we were teaching them, and so forth. All these methods of testing, rating, surveying, and interviewing would contribute to some measure of what we as teachers had done for our students. Such a project might yield some useful criterion, and then we could examine the correlation between this score and the student perception ratings we have been using. This correlation would show us how valid our instrument has been.

It seems intuitively likely to me that students would tend to perceive "good" teaching or "excellent" teaching, and therefore give average ratings of 3.5 or higher on item #10 in our summative scale, with the teachers who would be most likely to score high on whatever criterion emerged from the multi-trait-multi-method process I've described.