An Analysis of the Relative Importance Of Criteria Used On Student Evaluation of Teaching Effectiveness Instruments

Douglas Havelka (Email: havelkdj@muohio.edu), Miami University
Fred Beasley, (Email: beasley@nku.edu), Northern Kentucky University
Catherine S. Neal, Northern Kentucky University

ABSTRACT

Student evaluation of teacher effectiveness (SETE) has become commonplace as one measure of teaching performance in higher education. A study was performed to determine which criteria taken from several SETE instruments is considered more or less critical to learning by students. The data was gathered in the form of a magnitude measurement scale from students from multiple institutions with different missions and demographics. An analysis of the similarities and differences among the sample population, i.e. students, was performed and a discussion of the results is given.

INTRODUCTION

Student evaluation of teacher effectiveness (SETE) has become commonplace as one measure of teaching performance in higher education. Research has shown that institutions are using SETE for promotion and tenure decisions (summative purposes) as well as using it as a tool for improving teaching effectiveness and the quality of the learning experience (formative purposes). A study was performed to determine what criteria on a common SETE instrument are considered more or less critical to learning by students from multiple institutions of higher education. The data was gathered using the items from several SETE instruments in the form of a magnitude measurement scale. Students, and a small sample of parents of students, from multiple institutions with different missions and demographics were surveyed. A statistical analysis of some of the differences among the sample population, i.e. students and parents, was performed as well as a comparison of the institutions of the respondents. The details of the study, the outcomes of the data collection, the results of the statistical analysis, a discussion of the results and implications for various stakeholders, and suggestions for further research are presented below.

BACKGROUND

Student evaluation of teaching effectiveness (SETE) is often the most influential measure of performance used in promotion and tenure decisions at colleges and universities focused on teaching (Emery, Kramer et al. 2003). SETE is also used by students to select instructors prior to taking a course (Lewin 2003). SETE has been criticized for many reasons including: it is insufficient for measuring teacher performance(Sproule 2002), it is a disincentive to introducing rigor in the classroom (Millea and Grimes 2002; Emery, Kramer et al. 2003), it leads to grade inflation (Eiszler 2002), students' perceptions are inaccurate(Trinkaus 2002), and it is unrelated to student learning (Greimel-Fuhrmann and Geyer 2003).

In addition, research has found that students' evaluations are primarily based on teacher behavior, but also are affected by the students' liking or affection for the teacher or the teacher's charisma (Shevlin, Banyard et al. 2000) and affected by students' interest in the subject matter (Greimel-Fuhrmann and Geyer 2003). Some prior research has investigated factors that influence teaching effectiveness and learning, and students perceptions of these factors. Hill et al. found that the quality of the lecturer and the student support systems were most influential for quality education

(Hill, Lomas et al. 2003). And while some research has focused on the differences or relationship between students and faculty perceptions of SETE (Read, Rama et al. 2001; Sojka, Gupta et al. 2002), none were found that focused on analyzing the criteria contained in the SETE instruments and the perceptions of the importance of these items.

RESEARCH METHOD

To evaluate the relative importance of the criteria used on SETE instruments a magnitude measurement scaling technique was selected. The magnitude measurement scaling technique allows subjects to comparatively judge items on a ratio scale level. This gives more information about the relative importance of items when the items are being compared one to another. An extensive discussion of validation and reliability tests conducted on magnitude measurement can be found elsewhere (Howard 1981) and the use of magnitude measurement is widespread in multiple disciplines (Howard and Nikolai 1983; Havelka, Sutton et al. 1998). In addition, the magnitude measurement scaling technique is relatively easy to understand and apply. To establish the relative importance of each item, one item is selected as a benchmark. The item selected should be easy to understand and should be expected to have average importance. The selection of this benchmark is normally done by using the average ranked item from a small pilot study of students. This item is assigned a value of 100. All the remaining items are evaluated in relation to this item. Thus an item that is considered twice as important would be rated as 200, an item considered half as important rated at 50, and so forth. The geometric mean of all the individual ratings is then used as the point estimate for each item.

The magnitude measurement scaling technique was used to allow respondents to comparatively judge the relative importance of common SETE criteria. A benchmark item ("the instructor's attitude toward students is positive") was assigned a rating of 100 based on the results of a small pilot study. Each of the other items is rated on its importance to teaching effectiveness relative to this benchmark item. If an item is believed to be more (less) important than the benchmark, then the respondent must decide how much (less) important and choose a number to reflect the importance. Respondents were given both verbal instructions as well as a written instruction form that described the magnitude measurement scale.

The survey instrument used contained 22 items that were found in student evaluations of college teacher effectiveness (see Table 2 for a list of these items). All of the 22 items pertain to instructor-related aspects of teaching effectiveness. Items that were primarily course-related or subject matter items were removed for this study.

To analyze the data, the geometric mean of each factor is calculated and used as a point estimate for each factor. The geometric mean is used instead of an arithmetic mean due to the proportional nature of the data generated by the magnitude measurement technique. The geometric mean is calculated by taking the common logarithm (base 10) of each response, calculating the arithmetic mean of the logarithms, and then transforming to the geometric mean by calculating the antilog of the arithmetic mean of the logs(Snedecor and Cochran 1980). This geometric mean is then used as a point estimate of the overall rating assigned by students for each item. This data was then used to perform statistical analysis comparing the ratings by different groups of students and parents.

Data Collection

The magnitude measurement instrument was administered to a sample of 620 students from three mid-western universities. The majority of the respondents were students in multiple sections of information systems, marketing, and business law classes in the Fall of 2003. A profile of the students is found in Table 1.

Table 1 - Profile of Student Sample

Gender	Male	53%
	Female	47%
Class Rank	Freshman	2%
	Sophomore	50%
	Junior	27%
	Senior	18%
	Graduate	3%
Age	Median	20
GPA	Median	3.3

RESULTS

The geometric mean of the ratings for the 22 SETE items are presented in Table 2 (from most to least important. The item rated most important, on average, to the students for teaching effectiveness was that the teacher be knowledgeable about the subject matter that he/she is teaching. Other important items are that the teacher be able to clearly explain subject matter, and be fair and impartial in evaluating the performance of students. The items rated the least important to teaching effectiveness by the student sample concerned the ability of the instructor to promptly return exams and assignments, outline the course content in the syllabus, and explain the goals of the course.

Table 2 - Geometric Means of SETE Criteria Ratings

Is knowledgeable about the subject matter	186.08
Explains material clearly	168.38
Is fair/impartial evaluating work	156.31
Is helpful/responsive to questions in class	150.31
Is well-prepared for class	144.11
Shows interest and enthusiasm	141.91
Gives assignments/exams consistent with stated objectives	140.83
Is excellent	139.86
Is dedicated to high quality instruction	135.80
Deals with questions effectively	128.80
Is available outside of class	126.97
Is helpful outside of class	121.45
Effectively challenges me to think	119.45
Gives useful assignments	118.36
Holds students to high academic standards	108.67
Clearly communicates performance expectations and how measured	108.04
Manages class time effectively	104.16
Clearly communicates/follows course procedures	100.14
Attitude toward students is positive	**100**
Gives examinations that are challenging	98.27
Clearly explains goals/objectives of course	96.63
Outlines the course content in the syllabus	94.67
Promptly returns exams and assignments	93.99

STATISTICAL ANALYSIS

The first statistical analysis performed on the data was done to test the agreement among all the students toward the SETE criteria. Kendall's coefficient of concordance is used to determine the level of agreement among the ratings of the SETE criteria by the students. The Kendall's coefficient of concordance is calculated on the ranks of scores for all students to determine if agreement exists among the students regarding their ratings of the SETE criteria.

Larger values for Kendall's coefficient of concordance indicate stronger support for rejecting the hypotheses. Kendall's W ranges from 0 to 1, and a score near 0 indicates a lack of agreement among respondents. For the sample of students Kendall's W was found to be .184 with $p < .000$, indicating support for rejecting the null hypothesis.

Therefore, support exists that there is overall agreement among the students in their ranking of the items related to teaching effectiveness.

There were few significant differences in the ratings of the criteria by male and female students (see Table 3). Compared to men, women believed it was more important that instructors be fair and impartial in evaluating student work (t=2.198, df=609, p<.028), and believed it was more important that instructors manage class time effectively (t=2.709, df=602, p<.007).

To further investigate the ratings of the individual SETE criteria, the student sample was divided to analyze the data by gender, GPA, and class ranking. To determine which particular items students did not agree on, a comparison of each item using a t-test of differences in means is performed using gender, self-reported GPA, and class rank as student groupings. This analysis is performed using the means of common logarithms (base 10) of the magnitude data to minimize the effect on the error term due to the proportional nature of the data.

Students in the sample were divided into two segments based on whether their GPA was above or below the median GPA of 3.3. An analysis of differences in the responses of these groups found only that it is more important to students with lower GPAs that the instructor give assignments and exams that are consistent with course objectives (t=1.980, df=575, p<.048).

Students were also divided into underclassmen (freshman and sophomores) and upperclassmen (juniors, seniors, and graduate students). Compared to upperclassmen, it was more important for underclassmen that instructors were available outside of class (t=2.419, df=602, p<.016). For upperclassmen, it was more important that instructors effectively challenge students to think (t=3.245, df=609, p<.001), hold students to high academic standards (t=2.043, df=606, p<.041), and be fair and impartial in evaluating student work (t=3. 982, df=609, p<.000). These results are summarized in Table 3.

Table 3 - Significant Differences in Ratings of SETE Criteria By Gender, GPA, and Class Standing (t-test of differences in means, p<.05)

	Gender	Class	GPA
Manage class time effectively	*		
Is fair/impartial evaluating work	*	*	
Is available outside of class		*	
Effectively challenges me to think		*	
Holds students to high academic standards		*	
Gives assignments/exams consistent with stated objectives			*

A small sample of 40 parents of students was also obtained. An examination of the responses of students and parents of students revealed significant difference on five of the 22 items. For all five of these items, parents felt that the factors were more important than did students. Specifically, parents placed greater importance on the ability of the instructor to deal with questions effectively (t=2.702, df=656, p<.007), challenge students to think (t=3.487, df=658, p<.001), hold students to high academic standards (t=2.866, df=655, p<.004), clearly communicate course procedures (t=2.649, df=658, p<.008), and clearly explain course goals (t=2.232, df=653, p<.03). Table 4 presents a comparison of the item rankings for parents and students and indicates in bold the items where significant differences in the ratings of items were found. Perhaps the most striking differences in the rankings are found in the two items: "the instructor is helpful outside of class" and "the instructor effectively challenges me to think." Parents of students place much greater importance on the ability of professors to challenge their children in the classroom, while students care more about the ability of professors to help them outside of the classroom.

Table 4 - Student and Parent Rankings of SETE Criteria (1 = most important)

	Student	Parent
Is knowledgeable about the subject matter	1	1
Explains material clearly	2	2
Is fair/impartial evaluating work	3	6
Is helpful/responsive to questions in class	4	3
Is well-prepared for class	5	9
Shows interest and enthusiasm	6	8
Gives exams consistent with stated objectives	7	11
Is excellent	8	15
Is dedicated to high quality instruction	9	7
Deals with questions effectively*	10	5
Is available outside of class	11	12
Is helpful outside of class	12	20
Effectively challenges me to think*	13	4
Gives useful assignments	14	19
Holds students to high academic standards*	15	10
Clearly communicates performance expectations and how measured	16	13
Manages class time effectively	17	17
Clearly communicates/follows course procedures*	18	14
Gives examinations that are challenging	19	21
Clearly explains goals/objectives of course*	20	16
Outlines the course content in the syllabus	21	18
Promptly returns exams and assignments	22	22

The last area of statistical analysis was to investigate whether there were differences in rating the SETE items between students that attended different types of institutions of higher education. An analysis comparing the responses of the students from three schools was performed. Some significant differences were found. School 1 (S1) is an open admissions, regional university that offers four-year degrees and some masters degrees and is located in a suburban neighborhood of a large city. The student body consists of a significant number of students that are part-time, non-traditional, first generation, or commuter students that work full-time while taking classes. School 2 (S2) is an open admissions, two-year regional campus (of a large urban university) serving a primarily rural (even applachian) population consisting of a significant number of first generation, non-traditional, part-time, commuter students. School 3 (S3) is the main campus of a large state university (15,000 students) that offers undergraduate and some graduate degrees. It has a selective admissions process and is a rural, residential campus composed of primarily traditional undergraduate students. While there were no significant differences in the responses of students from S1 compared to students from S2, the S3 students did evaluate the importance of some items significantly differently than their counterparts. In all instances, the S3 students' ratings were lower than the ratings of the students from S1 and S2. Table 5 presents the items that where significant differences were found for S1 and S2 versus S3 (p-value < .05).

Table 5 - Differences in the Ratings of SETE Criteria --- S3 vs. S1 and S2

	S1	S2
Is helpful/responsive to questions in class	*	
Shows interest and enthusiasm	*	
Deals with questions effectively	*	
Effectively challenges me to think	*	
Clearly communicates performance expectations and how measured	*	
Clearly communicates/follows course procedures	*	
Outlines the course content in the syllabus	*	
Clearly explains goals/objectives of course	*	*
Promptly returns exams and assignments	*	*

DISCUSSION

The study finds that students believe that the most important characteristics of an effective teacher are that, 1) the teacher be knowledgeable about the subject matter that he/she is teaching, 2) the teacher be able to explain this subject matter clearly, and 3) the teacher be fair and impartial in evaluating student performance. These three characteristics of competency, communication skills, and fairness should be assessed in a good SETE instrument. It appears that the top two items, knowledge of the subject matter and explanation of the material are much more important than the remaining items. In fact, the top item, subject matter knowledge is considered twice as important as the lowest rated item, prompt return of exams and assignments. For institutions that have teaching as a primary mission, this may have implications for the skill set to look for when hiring instructors.

It should be noted that the benchmark item, attitude toward students is positive, was rated by the sample population of students much lower than the pilot study group (18^{th} out of 22 items). While this should not affect the outcomes of this study, it suggests that this item may not be the best choice as a benchmark for any future studies.

Also, there is some indication that the students rate items related to the instructor's ability to instruct more highly than the items related to class organization or management. For example, by using a rough classification scheme nearly all of the items (9 of 11) in the top half of the items are directly related to instruction while only 4 of 11 in the bottom half are. In contrast, 7 of the bottom 11 could be considered course management items and only one of the top 11. This could allow for some items to be combined or deleted from the instrument, if the length of the instrument is an issue. In fact, the results of this study may be used to help in paring down the length of the SETE instrument in general or a similar approach could be used at any institution to obtain information useful for that objective.

Based on the statistical analysis, there appears to be overall agreement among the students with regard to the relative importance of the items. The study did find some disagreement, however, among student sub-groups and some disagreement between students and their parents about the importance of specific items and thus the characteristics of an effective teacher. The significant differences between the male and female students (fair and impartial evaluating work and manage class time effectively) may be based on perceptions of the female students that there is (or was) gender bias in evaluation and so this would be a more important criteria for them.

Taking a broader view of the results of the statistical analysis reveals potential explanations for some of the significant differences. A plausible explanation for most of the differences is based on a general level of maturity, or even more specifically "who pays" for the education. It appears that the items where significant differences were found between the lower- and upper-class ranks are similar to the items where differences were found between parents and students and where differences were found among the students from different institutions. By taking a step back from the results, it seems that the differences among these sub-groups of students and the parents could be based on a common trait or factor, i.e. maturity. The results comparing the upper- and lower- class ranks reveals that the upper-class ranks rated the items "effectively challenges me to think" and "holds students to high academic standards" higher than did the lower-class ranked students. Similarly, the parents rated these items plus others higher than the students did. Also, the students from institution S1 rated several of the same items as the parents significantly higher than the students from S3. Comparing S1 to S3 reveals that a much higher percentage of the students at S1 are personally responsible for their education, i.e. they work or borrow money to pay for the classes. The students from S3, for the most part, receive family support or scholarship money for their education. It seems that the students from S1 are more similar to parents when it comes to rating the criteria. These observations suggest that there may be a common, underlying characteristic (mitigating co-factor?) that could explain the results. Unfortunately, we must leave this to future research to explore further.

College professors and administrators should be aware that there may be differences in the perceived importance of different teacher characteristics between male and female students, and between upperclassmen and underclassmen, and between students at different institutions. Moreover, universities should be aware that parents and students may have different perceptions as well. For example, parents are much more likely than their children to

believe that teachers should challenge students to think and hold students to high academic standards. Such a finding should be important to the marketing efforts of an institution.

In conclusion, while there is support from this study that students overall tend to agree which critieria on the SETE instruments are more and less important there are indications that some significant differences exist among sub-groups of students and between students and parents regarding the importance of certain criteria.

LIMITATIONS

One limitation of the study is that the sample consisted of only students majoring in business, and only students from three state universities in the midwestern U.S. It is not known if the results are generalizable to other types of college students and to students in other academic settings (e.g., arts and science or fine arts, private universities or international schools). The finding in this study of some differences in importance ratings among students of three different universities would suggest that some of the results at least may not be generalizable to other types of schools. Students from non-business-related disciplines could also be expected to have different perceptions about the importance of various teacher characteristics. Also, the results of the statistical analysis must be interpreted appropriately. Given the number of items on the magnitude measurement instrument to be evaluated and the sample size limits the usefulness of the pairwise comparison analysis of variance used in the study. Lastly, the survey given to students/parents only contained evaluation criteria related to the effectiveness of the teacher. Items that were primarily course-related or related to subject matter were removed for the study. Obviously, these items are important components of student satisfaction with a college course and should be part of a good SETE instrument.

SUGGESTIONS FOR FUTURE RESEARCH

As mentioned above, the study consisted of business students in three midwestern state universities. It would be interesting to examine if the same results are found in other academic settings. It seems intuitive that cultural, economic, and other differences found among students matriculating at private universities (as well as the parents of these students) and students outside the U.S. would lead to different perceptions about the relative importance of teacher characteristics. Future research should also examine the perceptions of teachers and administrators to see how their importance ratings differ from those of their students.

REFERENCES

1. Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher* Education 43(4): 483-502.
2. Emery, C. R., T. R. Kramer, et al. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education: An International Perspective* 11(1): 37-47.
3. Greimel-Fuhrmann, B. and A. Geyer (2003). Students' evaluation of teachers and instructional quality -- Analysis of relevant factors based on empirical evaluation research. *Assessment & Evaluation in Higher Education* 28(3): 229-238.
4. Havelka, D., S. G. Sutton, et al. (1998). A methodology for developing measurement criteria for assurance services: An application in information systems assurance. Auditing: *A Journal of Practice & Theor*y **17**(Supplement): 73-92.
5. Hill, Y., L. Lomas, et al. (2003). Students' perceptions of quality in higher education. *Quality Assurance in Education: An International Perspective 1*1(1): 15-21.
6. Howard, T. (1981). Attitude Measurement: Some Further Considerations. *Accounting Review* LVI(3): 613-621.
7. Howard, T. P. and L. A. Nikolai (1983). Attitude measurement and perceptions of accounting faculty publication outlets. *Accounting Review* LVIII(4): 765-776.
8. Lewin, T. (2003). New online guides allow college students to grade their professors. *New York Times. New York***:** A11.
9. Millea, M. and P. W. Grimes (2002). Grade expectations and student evaluation of teaching. *College Student Journal* 36(4): 582-591.

10. Read, W. J., D. V. Rama, et al. (2001). The relationship between student evaluations of teaching and faculty evaluations. *Journal of Business Educatio*n 76(4): 189-193.
11. Shevlin, M., P. Banyard, et al. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Educatio*n 25(4): 397-405.
12. Snedecor, G. and W. Cochran (1980). *Statistical Methods*. Ames, IA, The Iowa State University Press.
13. Sojka, J., A. K. Gupta, et al. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching* 50(2): 44-49.
14. Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review* 21(3): 287-295.
15. Trinkaus, J. (2002). Students' course and faculty evaluations: An informal look. *Psychological Reports* 91(3): 988.

NOTES