Developments In Business Simulation & Experiential Exercises, Volume 21, 1994

VALIDATING AN INSTRUMENT FOR STUDENT EVALUATION OF TEACHERS: SOME NOTEWORTHY BY-PRODUCTS

Kelly W. Crader, Clemson University
John K. Butler, Jr., Clemson University

ABSTRACT

We present strong evidence for the external validity, and some evidence for the construct validity, of a published instrument for student evaluations of teaching performance. For external validation, we replicated the instruments factor structure very closely, yielding five factors virtually identical to those of a previous study that used two different populations. For construct validation, we developed a preliminary theoretical model that predicts different effects of three sets of variables on student ratings of teachers. The model was partially supported. Student expectations about teachers had significant effects on all dimensions of teaching effectiveness. Class size and teacher experience had much weaker, and mostly negative, effects on the ratings. Student ratings of teaching effectiveness were driven more strongly by a student characteristic than by a teaching condition or a teacher characteristic.

INTRODUCTION

Student ratings of faculty teaching are often challenged on grounds that they are unethical or invalid. Since others have discussed ethical aspects, uses, and purposes of such ratings (Abrami, dApollonia & Cohen, 1990; Harari & Zedeck, 1973; Kerlinger. 1971; Kulick & McKeachie, 1975; Lester, 1982; LHommedieu, Menges, & Brinko, 1990; Marsh, 1984; Meredith, 1983; Wilson, 1982), we did not address those issues. Our study had two objectives, both of which addressed validity considerations. The first was to support the external validity of a questionnaire used for student evaluations of teachers. The second objective was to provide evidence for the construct validity of the questionnaire.

We investigated external validity in the context of a previous study by Wimberly, Faulkner, and Moxley (1978), who administered their questionnaire (the WFMQ) at a southeastern university. The courses were sociology, anthropology, and social work. The first year the researchers obtained 2,204 usable answer sheets; the second year, 2,152. For both years, factor analysis or the 40 items in the WFMQ yielded five factors: student development, teacher-task responsiveness, respects for students, teacher capability, and encouragement to students. This five-factor solution was judged the most interpretable for both sets of data. Our study replicated the WFMQ using data from four departments in a college of business.

Our approach to construct validation of the instrument relied on hypothesis testing. Construct validation is -. . . nothing more nor less than hypothesis testing," which requires predictions from theory (Hogan & Nicholson, 1988, p. 622). Abrami, et al (1990) noted the lack of theoretical argument in the literature on student ratings of teaching.

We synthesized a theoretical model that specified interrelationships among constructs relevant to student's evaluations of teaching performance. We then generated hypotheses from the model and used measures from the instrument to test the hypotheses. Assuming a valid model, support for the hypotheses would be evidence of construct validity of the instrument. Thus, the procedure was essentially the reverse of model testing, where one tests hypotheses assuming a valid instrument but a questionable theoretical model.

One type of model involves combining variables into blocks (Blalock, 1969). Each block is composed of conceptually similar variables, which are assumed to affect the dependent variables. The causal links that operate on the dependent variables from the independent variables, are the direct and indirect effects.

Fortunately, a model addressing student evaluations of teaching effectiveness has been specified and tested. Both the inventory of potential causes as well as the "blocking" of the causal variables have been accomplished by previous research. Anecdotal lore has identified a number of potential predictors of student ratings. However, contrary to the myths, empirical evidence supporting many "predictors" of teacher ratings is weak or inconsistent. Variables which most often fail to predict teacher ratings include: (a) students' grades and expected grades in a course; (b) the match between students' majors and the course subject matter; (c) students class rank; (d) the time of day a class meets; (e) the academic rank of the teacher; and (f) the number of years the teacher has been teaching (Abrami, et al 1990; Centra, 1979; Costin, Greenough, & Menges, 1971; Genova, Madoff, Chin, & Thomas, 1976; Grasha, 1977; Lester. 1982; McKeachie, 1979).

In contrast, four variables have been found to consistently predict teacher ratings: (a) student expectations about a course; (b) student expectations about a teacher, (c) whether a course is required or elective; and (d) class size (Centra. 1979; Costin, et al., 1971; Genova, et al., 1976; Grasha, 1977; McKeachie, 1979). Student ratings of teachers tend to be favorable when expectations about a course or a teacher are met, when a course is an elective, and when class size is small.

An implicit blocking has also occurred in that certain variables coincide with popular notions and others clearly contradict those notions. Kulik and McKeachie (1975) suggested three blocks of variables. They started with a demonstrated finding: student ratings of teaching effectiveness are not determined solely by the quality of teacher. They then outlined some common determinants of variation in student ratings. Finally, they blocked these determinants as follows, in order of importance: (a) student variables, including the student's general disposition toward the material, instructors, and instruction; (b) teaching conditions, including class size, substantive discipline, and subject matter within discipline; (c) teacher characteristics, including experience, academic rank, and research productivity. Abrami, et al (1990) identified the variables in these three blocks as "biasing" characteristics.

Although Kulik and McKeachie (1975) ranked their blocks of variables in order of empirically explained variation in student ratings of teachers, one can derive a theoretical rationale that is consistent with their review. Kulik and McKeachie (1975) noted that empirical evidence has indicated consistently that student ratings have had little or no effect upon teacher improvement. That is, student ratings do not seem to be very effective as guides for teacher development (LHommedieu, et al, 1990). A possible explanation for this apparent limitation of teacher ratings is that the process of assigning students to classes changes the composition of the students enrolled so that the feedback from any one class at any one time is minimally relevant for any other class at any other time. In effect, the composition of students enrolled might vary so drastically that the teacher is always one semester behind in meeting the needs, orientations, and skills of a particular batch of students.

Consistent with this possibility is the proposition that the relative importance of the predictors of student ratings of teachers is determined more by student characteristics than by teacher behaviors and abilities. More generally, Culbert and McDonough (1980) argued that evaluators tend to judge others' performance according to how well the performance is aligned with the evaluator's own values and expectations. The concept of alignment (together with the empirical findings concerning Kulik and McKeachie's three main-effect blocks, and the argument for potential effects due to rapid turnover) suggests a guiding principle; the relative importance of predictors of student ratings of teachers is related to the proximity of the predictors to student's perceptive domains.

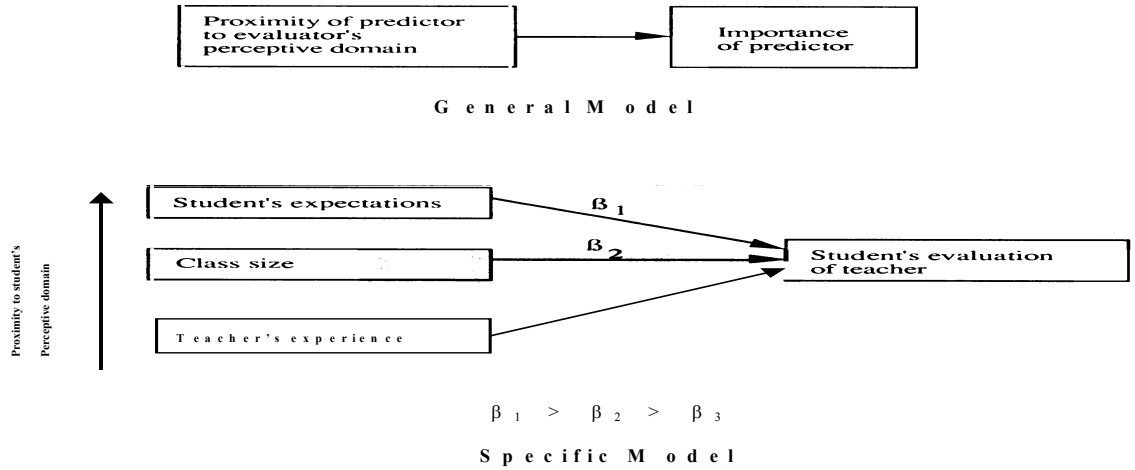
From this principle, we can hypothesize the same order of predictive power of the three main-effect blocks found by Kulik and McKeachie (1975). The block of predictors closest to the perceptions of students is student variables. Teaching conditions and teacher characteristics follow in that order. This argument suggests a conceptual model

composed of the three blocks of variables and their effects on student ratings of teachers. The block of student variables would have the strongest effect on student ratings. The effects of the teaching conditions and teacher characteristics blocks on students ratings would be weaker than that of the student variables block. Figure 1 portrays the above model, the basis for our construct validation of the WFMQ.

they were majoring in the subject area of the course, and (g) class rank. Students responded to all the questionnaire items on a five-point Liken-type format (strongly disagree to strongly agree).

Other data, also included for control purposes, were obtained from items on a cover sheet completed by the teacher of each section.

FIGURE 1
RELATIVE IMPORTANCES OF PREDICTORS OF A STUDENT'S EVALUATION OF TEACHING EFFECTIVENESS



METHOD

A southeastern university (different from Wimberlys) instituted a student-teacher evaluation system for all undergraduate courses. The system was based on a questionnaire that used only the 21 high-loading items of the WFMQ (Wimberly, et al., 1978). The other items were excluded to save class time in administering the questionnaire. The WFMQ was chosen because it focused on qualities of the teacher, not the course; and the primary purpose of the evaluation system was to be self-improvement of teachers. Secondarily, the instrument was intended as a performance appraisal tool, at the discretion of each teacher. The WFMQ was also unique in that some assessment of learning was included by requesting student's self-perceptions of their own learning. Seven items were added to the instrument with the intention of controlling for variables that had been found, at least occasionally, to affect student ratings of teachers. These items asked students: (a) it the course had been as they had originally expected, (b) if the teacher had been as they had originally expected, (C) if the course was required for them, (d) it they thought the ratings would be taken seriously, (e) if the form was being administered correctly, (f) if

These data were: college, discipline, class size, time section met, days section met, course level, the teacher's academic rank, the teachers years of teaching experience, and the teachers years of teaching the course under evaluation.

Prior to university-wide implementation, a pilot study was conducted in a college of business with four departments. The departments, the fraction of participating teachers, the fraction of participating course sections, and number of usable student responses were: Accounting and Finance (23/31 teachers, 45/84 sections, 1273 usable answer sheets); Economics (16/25 teachers, 36/62 sections, 1063 usable answer sheets); Management and Marketing (37/44 teachers, 62/113 sections, 2198 usable answer sheets); and Textiles (8/14 teachers, 10/21 sections, 217 answer sheets). Our data came from the 4751 usable answer sheets of the pilot study.

We tested the external validity of the WFMQ by replicating the study of Wimberly. et al (1978). First, we replicated the dimensionality of the items using confirmatory principal-axis factor analysis, then oblique

(promax) rotation of the factors. Squared multiple correlations were the initial communality estimates. In order to replicate the earlier study, we restricted the analysis to the 21 high-loading items and a five-factor solution was specified (Wimberly, et al., 1978).

We assessed the construct validity of the WFMQ by testing hypotheses generated by the model in Figure 1. We calculated measures of the five dimensions of teacher effectiveness by taking the means of items that loaded on a given factor (loading ≥ 30) and did not overlap on other factors. We then regressed these five scale means separately on three predictors of student-teacher ratings, each of which is included in one of the three blocks, student variables, teaching conditions, and teacher characteristics. Consistent with the literature review, we operationalized these predictors as: student expectations of the teacher, class size, and years of teaching experience respectively. We hypothesized that the above order would coincide with the variables' rank in explaining the student ratings of teachers.

RESULTS

The factor loadings of the 21 items are shown in Table 1. Rows denote the items and their factors that were predetermined by Wimberly, et al (1978). Columns represent the factors derived from our data.

The factor pattern matrix of Table 1 shows a remarkably close replication of the factors derived in two consecutive years by Wimberly, en al. (1978). The major exceptions are items originally allocated to "teacher capability" and "respect for students." One item ("not threatened by questions") loaded on "teacher capability" as expected, but with a weak loading. Another item ("permitted viewpoints other than own") loaded on "encouragement to students" rattier than "respect for students. Factor patterns are seldom so consistent; so seldom that factor scaling is often considered to be sample dependent (Armor, 1974). The close correspondence of factor patterns derived from three independent data sets, two from one institution and one from another, constitutes strong evidence of external validity of the WFMQ.

Table 2 shows the squared multiple correlations and standardized regression coefficients obtained from regressing each of the five teacher effectiveness measures on the three predictors. The amount of variance explained in the measures of teaching effectiveness is highly dependent upon what measure is under consideration. Certainly, the 27 percent of the variance explained in student development is noteworthy; the 10 percent explained in respect for students, much less so. In all instances, however, the explained variance is primarily a function of student expectations about the teacher, our indicator for student variable effects. Student

TABLE I
ITEM LOADINGS ON FIVE FACTORS USING PREDETERMINED DIMENSIONS

Item Number	Predetermined Factor/Item[a]	Derived factors I	II	III	IV	V
Student Development						
1	Gave new viewpoints	.67*	.00	.09	-.01	-.01
4	Stimulated subject interest	.92*	-.10	-.03	.06	-.01
5	Stimulated subject area [b]	.82*	-.06	-.04	.03	.02
8	Stimulated desire for learning	.71 *	.05	-.01	.12	.01
18	Improved understanding	.68*	.15	.05	.04	-.02
29	Improved interpretation	.60*	.15	.05	.04	-.02
Teacher-Task Responsiveness						
10	Explained expectations	.13	.58*	.05	.02	.02
15	Timely returned course work	-.06	.51 *	.06	-.07	.02
20	Made grading system clear	-.08	79*	.04	-.03	-.01
25	Gave clear progress reports	.06	.62*	-.10	.13	-.07
28	Gave fair evaluations	.23	43*	-.08	.05	.11
Teacher Capability						
17	Demonstrated subject command	.22	.11	55*	-.14	.02
22	Not threatened by questions [c]	.05	.07	.32*	.00	.24
26	Enjoyed teaching course	.02	-.01	.63*	.21	.01
30	Interested in subject	.02	-.04	77*	.11	-.05
Encouragement to Students						
3	Encouraged participation	.12	-.03	.06	53*	-.09
19	Encouraged student's best	.10	.23	.11	45*	-.02
21	Complimented students	-.03	.08	.00	.63*	.10
Respect for Students						
6	Respected students	.04	.04	.05	.21	.56*
13	Did not intimidate	-.02	.00	-.02	-.03	.85*
24	permitted other viewpoints [d]	.05	-.09	.06	,43*	.19
Reliability of Scale with * Items (Alpha)		.91	.82	.85	.80	.77

[a] All item descriptions arc paraphrased. A copy of the actual questionnaire may be obtained from the authors upon request

[b] Item 5 was dropped from "student development" because committee members could not agree on its meaning, i.e. content validity, and because of its high correlation with item 4 (r= .76).

[c] Item 22 was dropped because of low, two-dimensional factor loadings.

[d] Item 24 was re-assigned to the scale "encouragement to students," because it loaded high on factor IV and low on factor V.

expectations about teachers had strong, positive, and statistically significant effects on all five measures of teacher effectiveness. The strength of these effects, their consistency with findings of previous studies, and reasonably close alignment to the adopted model indicate support for construct validity of the measures.

DISCUSSION

The evidence for external validity of the questionnaire items developed by Wimberly, et al. (1978) is clear and consistent. Appropriate loading of items on predetermined factors, with but one exception, is

TABLE 2
REGRESSION OF TEACHER EFFECTIVENESS MEASURES ON SELECTED PREDICTORS: STANDARDIZED COEFFICIENTS

Effectiveness		Predictor of Teaching	Effectiveness	
Measure	R^2	Expectations for Teacher	Class Size	Teaching Experience
Student development	.27	.51***	~.03*	
Tchr-task rspnsvness	.23	43***	.05***	
Teacher capability	.17	.40***	.01	
Encouragement to stdnts	.18	.38***	~.08***	
Respect for students	.10	.30***	.08***	-.03

significant at p <.0001.

Note. All multiple correlations were
*p<,O5; **p<.01; ***p<A)01.

However, the other results portrayed in Table 2 were clearly inconsistent with what we hypothesized, particularly with respect to class size. The predictive power of class size (a teaching condition) and years of teaching experience (a teacher characteristic) are reversed from what we hypothesized. Years of teaching is a better predictor than class size, controlling for student expectations about the teacher. Although four of the five class size coefficients were significant, they were trivial in magnitude with their significance resulting from the large sample size. More disconcerting are the signs of the coefficients for class size. The signs of the coefficients are positive for teacher-task responsiveness and respect for students, opposite to the hypothesized negative relationships.

Selection effect could account for some of this reversal. To the extent that student selection of the "good" teachers (and avoidance of the "bad" teachers) influenced class size, there would tend to be a positive effect of class size on teacher ratings. An informal check on the teachers of the large classes (only four classes exceeded 60 students) revealed that the teachers were, in fact, popular. However, when the smallest and largest class size categories were excluded from the analysis, the results were still consistent with Table 2. Class size apparently had the following effects: (a) no effect on teacher capability scores, (b) a negative effect on student development scores and encouragement to students scores, and (c) a positive effect on teacher-task responsiveness and respect for students.

If selection effects were operating, we are left wondering why those effects were not mutually consistent and why they have not appeared in other studies. The effects of years of teaching experience on teacher ratings are all in the negative direction, a contradiction to other studies (Abrami, et al, 1990; Cohen, 1981). The magnitudes of the coefficients are also noteworthy, particularly with respect to teachers' task responsiveness and encouragement to students. Apparently, years of teaching experience contributes unfavorably to student perceptions of teaching effectiveness when the multidimensional nature of effectiveness is examined carefully.

remarkable, particularly since the contexts of administering the questionnaire were so different. There appear to be five dimensions of teacher effectiveness: student development, teacher-task responsiveness, teacher capability, encouragement to students, and respect for students.

The evidence for construct validity of the WFMQ through theory construction methods is not so remarkable but still provides some support for the instrument. Student expectations, our indicator for the student variable block, was more powerful in accounting for variation in teacher effectiveness than the indicators for teaching conditions and teacher characteristics. This was consistent with the model derived from Kulik and McKeachie (1975). The inability of class size to predict consistently the teacher effectiveness measures in the same direction is a challenge to the construct validity of the instrument. In almost all other studies, class size has been found to be negatively associated with student ratings of teachers. In our study, class size was not related to one dimension, negatively related to two dimensions, and positively related to another two.

One possible resolution to this inconsistency lies in the tendency of other studies to combine measures into global composites. When dimensional configurations are not explicitly built into teacher ratings, relationships between teacher ratings and predictor variables might be obscured or even reversed. For example, if a student-rating instrument relied heavily on items that denoted student development and encouragement to students, the relationship of a summary score to class size would tend to be negative. The reverse would be true if a preponderance of items focused upon teacher task responsiveness and respect for students. One wonders what might have resulted from the multitude of studies dealing with teacher ratings if measurements and analyses had separated the dimensions rather than combining them into global composites.

Contrary to the model, the indicator for teacher characteristics, years of teaching experience proved to be a more powerful predictor than did the indicator for teaching conditions, class size. However, this

reversal does not constitute strong evidence for abandoning the model since the magnitudes of the differences were small. Note that the full model is relatively untested since it is a blocked model and we used only one indicator for each block.

Much of the reported empirical evidence, which we have characterized as contrary to the mythology of rating predictors, might constitute another kind of myth -- one generated by measurement imprecision. Our findings with respect to teaching experience might be another example of needed clarifications via more precise measurement and analysis. For example, in our study, measures of teaching effectiveness were found to be negatively related to teaching experience.

We acknowledge a note of caution. Asking students at the end of the term about expectations at the beginning may lead to tautological associations, which are spurious for purposes of theory construction. Rather than the degree of fulfilled expectations leading to student perceptions of teacher effectiveness, the reverse might also be true. That is, retrospective expectations might be generated in accordance with and at the same time as perceptions of effectiveness. The associations between our measure of student expectations and the measures of teacher effectiveness might then be likened to correlations of two measures of the same thing.

However, there are two arguments in defense of using retrospective expectations First, Dillman (1978) found that people can recall a cognition if there is a realistic time lag and a low requirement for cognitive effort. For the current study, the time lag was only 14 weeks and the task was cognitively very simple. Thus, it is not unreasonable to assume that students could recall what they originally had expected from the teacher and course.

The second defense of using retrospective expectations, at least in the current study, is that the correlation of two measures of the same variable suggests either concurrent validity (Kerlinger, 1967) or construct (convergent) validity (Campbell and Fiske, 1959) The strong effect of student expectations on teacher ratings supports either construct validity by supporting the model in Figure 1, or it supports concurrent validity. If fulfilled student expectations cause favorable ratings, construct validity is supported. If student expectations are generated simultaneously with student perceptions of teacher effectiveness, then concurrent validity is supported. We can not have it both ways, but either way there is evidence of validity

Most important, we emphasize that teachers' behaviors are (and should be) the focus of a student rating system. There is ample support, in the literature on attitudes, for the hypothesis that students' attitudes influence their behaviors in rating teachers and that teachers' behaviors influence students' attitudes (Ajzen and Fishbein, 1977). However, a common erroneous assumption underlying students' ratings of teachers is that teachers' behaviors are the only characteristics being assessed.

Another erroneous assumption is that teachers' behaviors indicate their abilities. The implied causal chain is that teachers' abilities cause their behaviors, their behaviors cause students' attitudes, and students' attitudes cause students' ratings Thus, teachers' abilities have only a twice-removed indirect effect on the ratings and the assumption that ratings indicate abilities can be particularly misleading. It can possibly be damaging when an administrator uses the ratings to make decisions about salary, promotion, or tenure.

Our data, in conjunction with the model, suggest that teachers' behaviors and abilities are not the only and probably not the most important variables affecting students' ratings of teachers. Practically speaking, support of the model indicates that teachers might have less control over their ratings than commonly believed. This is not to say that teachers are totally helpless, but that there are limits within which they can operate effectively. Those limits are, in part, determined by variables intervening between teachers' abilities and students' ratings of teachers. In fact, we are sufficiently convinced of the above-mentioned causal chain (teachers' abilities --> teachers' behaviors --> students' attitudes --> students' ratings) to suggest that no ~ effects should be proposed between teachers' abilities and students' ratings of teacher effectiveness. The WFMQ measure of teachers' effectiveness might have construct validity as a measure of teaching effectiveness, but it is not a measure of teachers' abilities.

REFERENCES

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990) Validity of student ratings of instruction: What we know and what we do not. Journal of Educational Psychology, fig, 219-231

Ajzen I. and M. Fishbein (1977) Attitude-behavior relations: a theoretical analysis and review of empirical research. Psychological Bulletin, ~4, 888-918.

Armor D. J. (1974). Theta reliability and factor scaling. In H. L. Costner (ed.), Sociological Methodology 1973-74. San Francisco: Jossey Bass.

Blalock, H. M. Jr. (1969) Theory construction Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Campbell, D. T., & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix Psychological Bulletin 55, 81-105

Centra, J. A. (1979) Determining faculty effectiveness. San Francisco: Jossey-Bass.

Cohen, P. A. (1981) Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, ii. 281-309.

Costin, F., Greenough, W. T., & Menges, R. J. (1971) Student ratings of college teaching: reliability, validity, and usefulness. Review Educational Research 41 511-535.

Culbert, S. A., & McDonough, J. J. (1980) The invisible war Pursuing self-interest at work. New York: Wiley.

Dillman, D. A. (1978) Mail and telephone surveys: The total design method New York: Wiley.

Genova, W. J., Madoff, M. K., Chin, R. & Thomas, G. B. (1976) Mutual benefit evaluation of faculty and administrators in higher education Cambridge, MA. Ballinger Publishing Co

Grasha, A. F. (1977) Assessing and developing faculty performance Principles and models. Cincinnati, Ohio: Communication and Education Associates.

Harari, 0. & Zedeck, 5. (1973). Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology 55, 261-265

Hogan, R., & Nicholson, R. A. (1988) The meaning of personality test scores. American Psychologist, fl. 621-626

Kerlinger, F. N. (1967) Foundations of behavioral research New York. Holt, Rinehart and Winston

Kerlinger, F. N. (1971) Student evaluation of university professors School and society, ~, 353-356.

Kulik, J. A. & McKeachie, W. J. (1975) The evaluation of teachers in higher education. In F, N. Kerlinger (ed.), Review of research in education Vol. 3 Itasca, NY: Peacock Publishers.

Lester, D. (1982) Students' evaluation of teaching and course performance Psychological Reports, 55, 1126

L'Hommedieu, R. L., Menges, R. J,, & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. Journal of Educational Psychology, 82, 232-241.

Marsh, H. W. (1984) Students' evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility Journal of Educational Psychology ZS 707-754

McKeachie, W. J. (1979) Student rating of faculty: a reprise. Academe .15 384-397

Meredith, G. M (1983) Diagnostic value of composite student-based ratings of instruction Psychological Reports, 5~, 549-550.

Wilson, T. C. (1982) Student-faculty evaluation forms in marketing. Journal of Marketing Education, 4 (Spring), 7-14.

Wimberly, R. C., Faulkner, G. L. & Moxley, R. L. (1978) Dimensions of teacher effectiveness Teaching Sociology, ~, 7-20.