Integrating Theoretical Perspectives Of Student Rating Behavior

Louise Hatfield, (E-mail: lohatf@wharf.ship.edu), Shippensburg University
Jonathan Kohn, (E-mail: jwkohn@wharf.ship.edu), Shippensburg University

Abstract

This paper examines attribution, grade-leniency/stringency, motivation, and construct validity theories of student-rating behavior. Previous research has tended to treat these theories as mutually exclusive. However, our findings suggest that these theoretical perspectives should be integrated into a unified theory of student-rating behavior. The R-squares found in this study are larger than in any previous research, which we attribute largely to the rigor of using structural equation modeling (SEM) to analyze the data. The significant bias that was found in this study suggests the need to rethink the use of student ratings when evaluating faculty performance.

Introduction

Student ratings of faculty are a controversial issue. Despite their shortcomings, universities have long relied on these ratings as a primary method of assessing professor teaching effectiveness. These shortcomings include bias and validity concerns in the survey instrument and rating process. One of the major bias issues is grade leniency, which suggests that students give high ratings in appreciation for high grades. However, this premise has been challenged and various theories have been offered as an explanation for the positive correlation between grades and ratings of faculty. This paper will examine each of these theories and evaluate their predicted effects on student ratings of professor effectiveness. Using AMOS software for structural equation modeling (SEM), a comprehensive model will be presented in an attempt to integrate the major premises of these theories, thereby suggesting that the theories are basically complementary rather than mutually exclusive.

Literature Review

Much research has focused on explaining the positive relationship between grades and professor effectiveness. The effects of motivation on professor ratings are probably the most agreed upon systematic influence in student ratings of faculty. It has been demonstrated that student motivation, represented by student interest and course type (elective/required), plays a significant role in student ratings of professor effectiveness (Howard & Maxwell, 1980; Hoyt 1973; Marsh, 1984; Marsh & Duncan, 1992). Howard and Maxwell (1980, 1982) modeled the relationship between student motivation, student learning, expected grades, and student satisfaction with the instructor and field of study. These and other studies show that motivation and learning are more highly correlated with ratings of professor effectiveness than is expected grade with professor effectiveness. The authors conclude that student motivation drives the correlation between grades and student satisfaction with the instructor. Therefore, the correlation between grades and ratings of professor effectiveness is an expected artifact, rather than an indication of a direct relationship between grades and ratings of professor effectiveness. Using path analysis, Marsh (1984) also concluded that prior subject interest had a stronger impact on student ratings of various professor effectiveness characteristics than did grades. Additionally, simple classifications (required versus elective) and expanded categories of course type have been found to be significantly correlated with ratings of professor effectiveness (Aleamoni, 1981; Centra, 1993; Feldman, 1978; Marsh & Dunkin, 1992).

Because of student motivation effects, Marsh and Dunkin (1992) express concern regarding the fairness of using student ratings to evaluate faculty. Student motivation is cited as a variable related to student ratings that requires control in order to promote fairness in comparing professors that teach interested students versus

uninterested students (Cashin, 1995). While past research clearly suggests student motivation has an impact on ratings of professor performance, the exact nature of this impact needs further investigation.

Construct validity theory proposes that student ratings reflect student learning and, therefore, measure professor teaching effectiveness. That is, higher student ratings for the instructor indicate greater student learning. Some studies have demonstrated that classes with the highest student ratings have also performed best on standardized final examinations in multi-section classes (Marsh & Roche, 1997). In addition to the earlier studies that provided the foundation for validity theory, numerous factor analytic studies have been conducted to investigate the validity of student ratings (Cashin 1988; Feldman 1989; Howard, Conway, & Maxwell 1985; Marsh 1984; Marsh & Duncan 1992).

Supporters of the validity theory argue that, despite the existence of many variables that might distort the ratings, bias has little overall impact (d'Appollina & Abrami 1997; Marsh & Roche 1997; McKeachie 1997). Others feel that these influences are strong enough to propose adjustments to the ratings, such as re-scaling (Greenwald & Gilmore, 1997a; Haladyna & Hess 1994). There seems to be widespread support for at least some degree of validity theory.

Using SEM, Greenwald and Gilmore (1997a&b) found support for grade leniency theory by suggesting that only grade leniency allows for a negative workload→grade relationship. This relationship is explained by students' willingness to work harder in order to avoid very low grades. This negative relationship between workload and grades has been observed in other studies (Marsh, 1980). However, other explanations have been offered for the negative relationship, such as subject difficulty and student capability (McKeachie, 1997).

The grade→professor effectiveness rating relationship has also been used to examine attribution effects. Research in this area has shown that individuals often credit themselves for success and blame others for their failure (Ross & Fletcher, 1985). Differences in ratings of professor effectiveness have been found when students were divided into groups based on whether the grade received was worse, the same, or better than expected, with the lowest ratings occurring in the lowest grade group found (Bridgeman, 1986; Owie, 1985).

Gigliotti and Buchtel (1990) found minimal attribution bias effects, regardless of unit of measure (within or between-class). The authors also found simultaneous external and internal attribution effects, rather than a clear-cut discriminant pattern. While they concluded that attribution would not affect the validity of student evaluations, they acknowledged that their model lacked the sophistication to accurately represent the complexity of the attribution construct. Although they suggest that attribution does not support grade-leniency effects, it does support grade-stringency effects where strict-grading instructors receive lower ratings. However, they found inconsistent support for both grade leniency and grade stringency. This lack of support may be due to the fact that expected grades recorded at the beginning of the semester are unrepresentative of realistic grade expectations, because students have not yet received any performance feedback.

Many early student-rating studies have been criticized for basing their conclusions on descriptive correlations, rather than on more rigorous statistical tests and procedures (Chacko, 1983; Holmes, 1972; Marsh, 1984; Powell, 1977; Stumpf & Freedman, 1979). Many studies have also been criticized for neglecting the multivariate nature of student ratings, utilizing inappropriate unit of analysis (within versus between-class), and not reporting effect sizes (Haladyna & Hess, 1994; Marsh & Roche, 1997). The utilization of different unit of analyses, as well as different statistical techniques and univariate analysis, may contribute to the variation in reported findings.

Hypotheses

Four theories have been offered in explanation of the positive relationship between grades and ratings of faculty (Greenwald & Gilmore, 1997a). Grade leniency and attribution suggest that grades directly affect ratings of faculty. Construct validity and motivation assert that a third variable (learning) positively affects both grades and ratings, thus, resulting in a positive relationship between grades and ratings. Hypotheses are developed for each of these theories.

Grade Leniency/Stringency This theory suggests praise induces liking for the individual giving the praise (Aronson & Linder, 1965; Hatfield & Kohn, 2003). In the context of student ratings, praise is interpreted to be high grades and liking is translated into high faculty ratings. Grade leniency theory suggests that there is a causal relationship between expected grades and ratings of faculty. Further, Greenwald and Gilmore (1997a) suggest that there is a negative relationship (grade stringency) between students working hard and expected grades. In courses that have strict-grading policies students have to work hard in order to avoid very low grades, yet overall grades are still lower than in classes with easy-grading professors. These premises suggest the following hypotheses:

H_1: *The higher the average expected grade, the higher the average professor effectiveness rating.*
H_2: *The higher the average student effort (worked harder), the lower the average expected grade.*

Construct Validity This theory suggests that high instructional quality induces high student learning, which results in higher grades and higher professor ratings (Cashin & Downey, 1992; Cohen, 1981; Feldman, 1976 & 1989; Marsh, 1984). Therefore, the following hypotheses are provided to evaluate construct validity:

H_3: *The higher the average student learning, the higher the average professor effectiveness rating.*
H_4: *The higher the average student learning, the higher the average expected grade.*
H_5: *The higher the average professor effectiveness rating, the higher the average student learning.*
H_6 *The higher the average worked hard rating, the higher the average student learning.*
H_7: *Professor effectiveness rating and expected grades are not causally related.*

Student Motivation This theory suggests that student motivation positively affects both grades and ratings of faculty, through student learning, thereby resulting in a positive correlation between grades and ratings of faculty (Aleamoni, 1981; Braskamp & Ory, 1994; Centra, 1993; Kohn & Hatfield, 2001; Marsh, 1984; Marsh & Dunkin, 1992). Student motivation results in more student learning and appreciation for the course and instructor, which leads to higher grades and higher professor effectiveness ratings. Researchers have identified two measures of student motivation: course-specific and general (Howard & Maxwell, 1980; Marsh, 1984). These indicators of student motivation will be examined in this study—student interest in the subject matter of the rated course and course type (major or elective, versus required or core course). Student interest is a course-specific measure, whereas course type is a general measure. The following hypotheses are designed to test the impact of student motivation in student rating behavior:

H_8: *The higher the average student interest, the higher the average student learning.*
H_9: *Lack of choice in course (required or core courses) results in lower average student learning.*
H_{10} *The higher the average student learning, the higher the average expected grade.*
H_{11}: *The higher the average student learning, the higher the average professor effectiveness rating.*
H_{12}: *Professor effectiveness rating and expected grades are not causally related.*

Attribution Attribution theory suggests that people take credit for desired outcomes of their actions and blame others for undesired outcomes (Davis & Stephan, 1980; Gigliotti & Buchtel, 1990; Hatfield & Kohn, 2003; Ross & Fletcher, 1985; Simon & Feather, 1973). In the context of student ratings, students attribute high grades to their own ability and low grades to poor instruction. Therefore, there should be only a modest relationship between expected grades and professor effectiveness ratings for students receiving high grades. However, the grade-rating relationship should be identifiable for students receiving low grades. These premises suggest the following hypotheses:

H_{13}: *For classes with high average expected grades, average expected grade will be independent of average professor effectiveness rating.*
H_{14}: *For classes with low average expected grades, the lower the average expected grade the lower the average professor effectiveness rating.*

Research Methods

The student rating survey contained eight items, which students rated on a six-point Likert scale: (1) strongly agree, (2) agree, (3), slightly agree, (4) slightly disagree, (5) disagree, (6) strongly disagree. The first six items were designed to examine professor effectiveness, with the sixth item being a global item. Student learning

was assessed by item 7 and course specific student interest by item 8.

1. The course requirements, including grading system, were explained at the beginning of the semester.
2. The professor provides feedback on exams and assignments.
3. The professor is willing to answer questions and assist students upon request.
4. The professor uses examples and practical applications in class, which aid in my understanding of the material.
5. The professor encourages students to analyze, interpret, and apply concepts.
6. The professor was effective teaching this course.
7. I learned a significant amount in this course
8. I am interested in the subject matter of this course.

These items were selected based on past research, which suggests the desirability of global items that address professor effectiveness (#6) and student learning (#7) factors, and the need to control for student interest (#8) (Cashin, 1995). Items one thru five address commonly used dimensions of professor effectiveness in student rating research (Braskamp & Ory, 1994; Cashin, 1995; Centra, 1993; Feldman, 1989; Marsh, 1991).

Students completed a student data sheet that contained demographic items, two grade-related items, and one general student motivation item: (1) The grade I expect to achieve in this course, (2) I worked harder in this course than in most of my other courses, and (3) course type. All response options were designed so that students could use opscan sheets to report their ratings. The scale for expected grade was: 1.A, 2.B, 3.C, 4.D, 5.F. The agree-disagree Likert scale noted above was also used for the 'worked harder' item. Five categories of course type were provided: A. required by major/minor, B. elective in major/minor, C. general education requirement, D. free elective, E. program core course. These items reflect commonly used measures in testing for grade leniency and motivation effects on student ratings of faculty (Greenwald & Gilmore, 1997a&b; Howard & Maxwell, 1980; Marsh, 1984).

Sample and Procedures

Data were collected from students and professors in the three colleges (business, arts and science, and education) at Shippensburg University at the end of the first semester of the 1997-1998 academic year. Classes were included in the sample from professors volunteering and by request (in order to insure adequate representation from all colleges and departments), a mix of student classes (such as freshman and senior), and a mix of professor characteristics (such as gender, race, degree, and rank).

Nine hundred and forty-one students and 45 professors were included in the sample, with the largest percentage (51) of faculty in Arts and Sciences, and equal percentages in Business and Education. The largest percentage (36) of students were seniors, followed by sophomores at 19 percent, juniors at 18 percent, freshmen at 14 percent, and graduate at 13 percent.

Variables and Measures

The professor effectiveness dependent variable is a composite measure, developed by averaging the ratings of the six professor effectiveness items. The reliability coefficient, alpha, for the composite professor effectiveness measure is 0.84. *Expected Grade* is both a dependent and independent variable, and is used directly as reported. *Student Learning, Student Interest,* and *Worked Hard* are also used as directly reported in the survey instrument. The *Course Type* categories were collapsed into a two-category independent variable: (1) major/minor/elective, and (2) required/core course. There are two measures of student motivation: *Course Type* and "*Student Interest* in the subject matter of this course". There are two measures of grade leniency: *Expected Grade* and "*Worked harder* in this course than in most other courses". *Student Learning*, a self-reported rating, is a construct validity measure.

The scales for five variables (professor effectiveness, expected grade, student learning, student interest, worked hard) were reversed so that interpreting the findings would be more consistent with the way these variables

are typically referred to, e.g., low to high. For example, the higher the student learning rating, the more the student learned, etc. The course type variable is categorical, and, thus did not need to be reversed.

Rather than using the individual student data (within-class data), class averages were calculated for between-class analyses. The between-class unit of analysis is recommended for examining student-rating data (Haladyna & Hess, 1994; Marsh & Roche, 1997). The between-class analysis minimizes extraneous individual variances and biases. Further, comparisons of grade-bias studies using the two types of unit of analysis reveals a larger effect size in the between-class group (Feldman, 1976; Stumpf & Freedman, 1979).

In order to test the attribution hypotheses, the sample was split by average expected grade, yielding a high-expected grade group and a low-expected grade group. The dividing point was chosen by percentile in order to assure almost equal numbers in the high and low expected grade groups. The high-expected grade group consists of twenty-two classes with average expected grades from 4.27 to 5.00, and the low-expected grade group consists of twenty-three classes with average expected grades from 3.22 to 4.20.

Analysis and Results

Descriptive statistics (means, standard deviations, and correlations) for all the variables used in this study are presented in Table 1. The hypotheses will be tested on the between-class data, using regression analysis and structural equation modeling (SEM). While there are many goodness-of-fit statistics in SEM, this study will report three of the most popular measures (CFI, NFI, Chi-square/df), with Comparative Fit Index (CFI) being the primary fit-statistic used in this study. [1] Path coefficients are tested for significance using Critical Ratios (CR). A CR of approximately 2.00 is considered to be statistically significant at the .05 level (Bollen, 1989).

[1] A 1.0 CFI or NFI suggests a perfect fit and if under .9 the model can probably be improved (Bentler and Bonnett, 1980). Chi-square/df ratios of up to 3 are indicative of acceptable fit models (Marsh and Hocevar, 1985). CFI is less affected by sample size than is NFI or the Chi-square ratio (Kline, 1998).

Table 1 Between-Class Means, Standard Deviations, and Correlation Coefficients Whole Sample

Variable	Mean	s.d.	1	2	3	4	5
1. Prof Effectiveness	5.36	.38					
2. Expected Grade	4.22	.45	.56**				
3. Student Learning	5.07	.47	.69**	.57**			
4. Student Interest	4.85	.70	.49**	.52**	.77**		
5. Worked Hard	4.13	.78	.12	-.24	.23-	.01	
6. Course Type[a]			-.32*	-.10-	.31*	-.29	-.36*

N=45; *p<.05; **p<.01

Low Expected Grade Group

Variable	Mean	s.d.	1	2	3	4	5
1. Prof Effectiveness	5.21	.38					
2. Expected Grade	3.84	.24	.68**				
3. Student Learning	4.89	.44	.73**	.71**			
4. Student Interest	4.55	.74	.48*	.42*	.74**		
5. Worked Hard	4.33	.67	.28	.04	.16	.07	
6. Course Type[a]			-.42*	-.00 -	.37	-.35	-.39

N=23*p<.05; **p<.01

High Expected Grade Group

Variable	Mean	s.d.	1	2	3	4	5
1. Prof Effectiveness	5.50	.32					
2. Expected Grade	4.62	.22	.23				
3. Student Learning	5.26	.43	.50*	.19			
4. Student Interest	5.17	.51	.21	.10	.73**		
5. Worked Hard	3.93	.85	.23	-.10	.58**	.20	
6. Course Type[a]			-.15	-.10-	.22	-.15	-.42

N=22; *p<.05; **p<.01

[a]A mean and standard deviation are not listed for course type because it is a categorical variable. The sign of course type is dependent on how it is coded. The correlations are negative here because required and core program courses were coded as 2's and major, minor, and elective courses were coded as 1's.

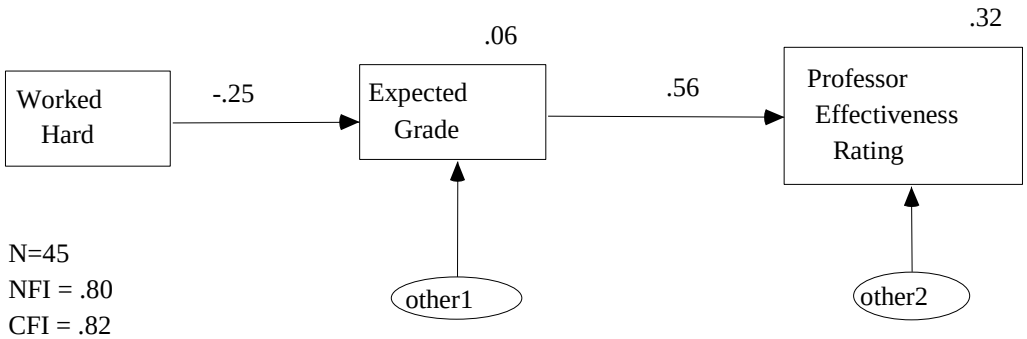
In addition, the Akaike Information Criteria (AIC) will be used to compare nonhierarchical models, with the lowest AIC indicating the preferred model (Kline, 1998). When using the Chi-square difference test, if the difference is not significant then the more restricted model is preferred on the principle of parsimony (Anderson & Gerbing, 1988).

Grade Leniency/Stringency Hypotheses 1 and 2 were tested using SEM. The two relationships predicted by these hypotheses were used to construct a SEM model using Amos software (see Figure 1).

H_1: The higher the average expected grade, the higher the average professor effectiveness rating.
H_2: The higher the average student effort (worked harder), the lower the average expected grade.

H1 is supported, as the relationship is positive and significant at the .05 level. H2 is negative as predicted, however, the linkage is not significant at the .05 level. Thus, H2 is not supported. The model fit was not particularly strong (CFI = .82.), suggesting there are missing relationships.

Figure 1
Grade Leniency/Stringency



N=45
NFI = .80
CFI = .82
Chi-squared = 4.81
df = 1
Chi-square/df = 4.81

The Construct Validity Hypotheses 3 through 7 were used to construct a SEM model, which is reported in Figure 2.

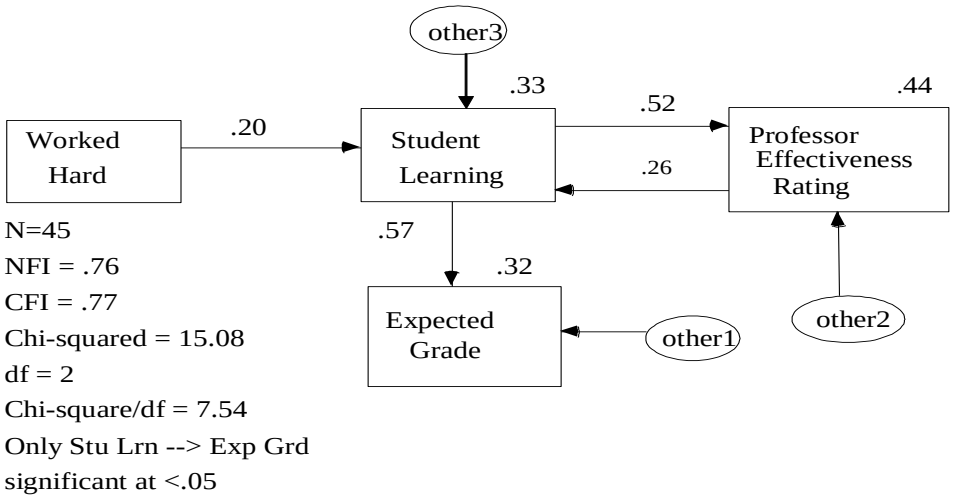
H_3: The higher the average student learning, the higher the average professor effectiveness rating.
H_4: The higher the average student learning, the higher the average expected grade.
H_5: The higher the average professor effectiveness rating, the higher the average student learning.
H_6: The higher the average worked hard rating, the higher the average student learning.
H_7: Professor effectiveness rating and expected grades are not causally related.

The results provide little support for construct validity, as the analysis only supports one of the 5 hypotheses. H4 (Student Learning → Expected Grade) is supported, as the relationship is positive and significant at the .05 level. H3, H5, and H6 are all positive, but not significant at the .05 level. Thus, Hypotheses 3, 5, and 6 are not supported. To test H7 we investigated the modification indices, which show relationships that, statistically, should be added to the model. These indices suggest there is a direct relationship between expected grades and professor effectiveness. Thus, H7 was not supported. Additionally, the model fit is poor (CFI = .77) and, therefore, should not be considered a useful measure of strength of relationship.

Motivation hypotheses 8 –12 are also tested using SEM. This analysis is reported in Figure 3.

H_8: The higher the average student interest, the higher the average student learning.
H_9: Lack of choice in course (required or core courses) results in lower average student learning.
H_{10}: The higher the average student learning, the higher the average expected grade.
H_{11}: The higher the average student learning, the higher the average professor effectiveness rating.
H_{12}: Professor effectiveness rating and expected grades are not causally related.

Figure 2
Construct Validity



The model fits the data quite well (CFI = .91), and provides support for Hypotheses 8, 10, and 11, as the predicted relationships are significant at the .05 level. H9 is negative as predicted, but is not significant at the .05 level. H12 was tested via the modification indices. These indices suggest there is a direct relationship between expected grades and professor effectiveness. Thus, H12 is not supported.

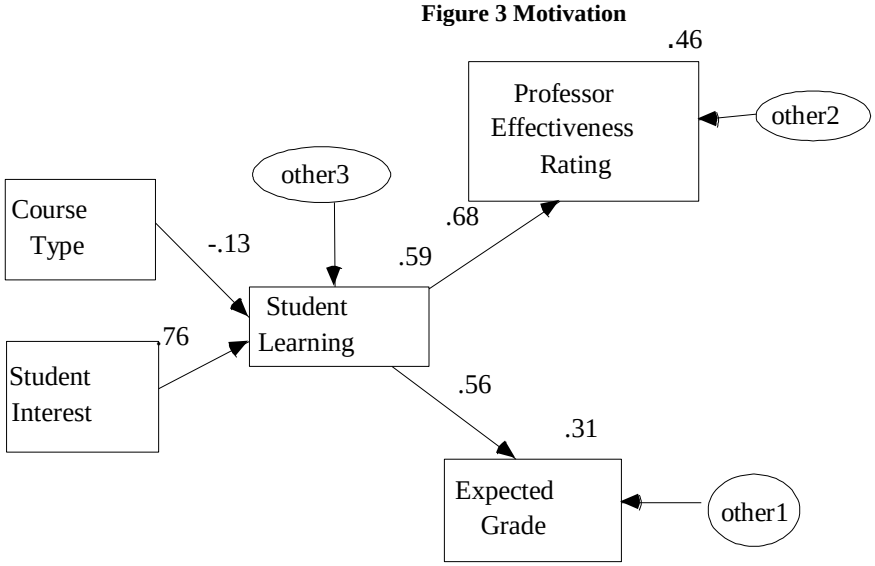
Attribution Hypotheses 13 and 14 were tested using stepwise regression analysis.

H_{13}: *For classes with high average expected grades, average expected grade will be independent of average professor effectiveness rating.*

H_{14}: *For classes with low average expected grades, the lower the average expected grade the lower the average professor effectiveness rating.*

The sample was split into a high-expected grade group and a low-expected grade group. A regression model was developed for each group assigning the Professor Effectiveness Rating as the dependent variable and Expected Grade as the independent variable. This analysis supported both hypotheses. That is, Expected Grade did not explain a significant amount of variance in Professor Effectiveness (R^2 = .05, F = 1.08). In the low expected grade group, Expected Grade explained 46 percent of the variance in Professor Effectiveness Rating, and the F statistic was significant at the .000 level with a positive direction of influence.

Figure 3 Motivation



N=45
NFI = .86
CFI = .91
AIC = 32.32
Chi-square = 14.32
df = 6
Chi-square/df = 2.39
Crs Typ --> Stu Lrn not significant
All other paths significant < .05

Table 2 Regression Analysis for Professor Effectiveness Using the Expected Grade Independent Variable (Attribution Theory)

	High Expected Grade Model[a]	Low Expected Grade Model
R	.23	.68***
R^2	.051	.46
Adjusted R^2	.004	.44
F	1.079	17.91***

Ṅ=45
***p<.0001
[a]Expected Grade did not enter the model at p<.05

The above analysis illustrates that each theory, by itself, either failed to explain student rating behavior or did so only partially.

New Perspectives on Student Ratings

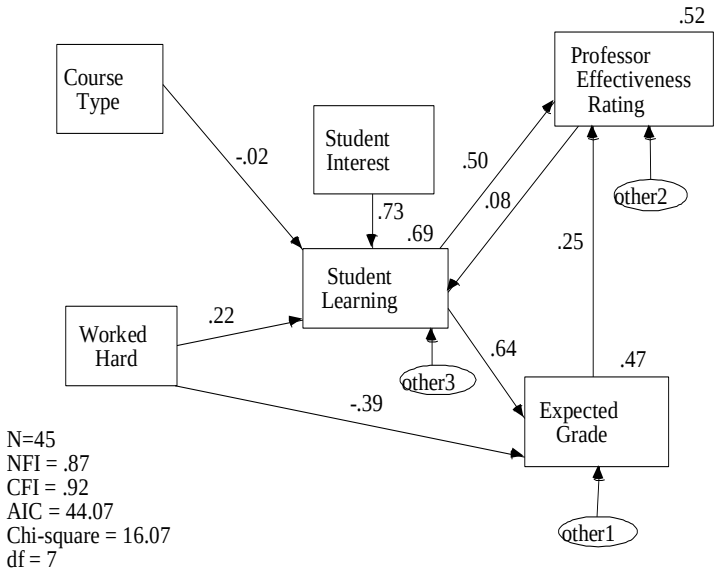
One of the problems with testing each of the above theories in isolation of each other is that intervening and moderating effects on the predicted relationships are not taken into account. Such effects may suppress or reinforce the predicted relationships. Thus, to accurately assess the presence of the theorized relationships, all the variables of interest need to be included in the same model. This section will integrate the findings predicted by the various

theories, using structural equation modeling.

All of the predicted direct relationships proposed in the grade leniency/stringency, construct validity, and motivation theories were used to construct an integrated structural model. Analysis of the model reveals that it fit the data well (CFI = .92 AIC = 44.07), and resulted in a R^2 of .52 for professor effectiveness. This model is presented in Figure 4 as Model A. However, several path coefficients were not significant. Removal of these paths was evaluated by iteratively removing the path with the lowest critical ratio (CR), rerunning the model with the path deleted, and then inspecting the CRs of the remaining paths. As a result of this procedure, two of the three paths were removed, namely, Professor Effectiveness Rating→Student Learning (Construct Validity), and Course Type→Student Learning (Motivation) were deleted. All remaining paths were significant. Inspection of the modification indices of the final model suggested that a linkage should be added between Course Type and Professor Effectiveness Rating. Insertion of this path is supported by a more complex motivation theory, supporting both the direct effect and indirect effect of student motivation on professor effectiveness. These results are presented in Model B. All paths are significant, the fit is very good (CFI = .96, AIC = 38.73), and the R^2 is .52. The R^2 of the integrated model is substantially higher than any of the individual models reported in Figures 1, 2 and 3. This model contains all the original variables and reveals the intricate nature of student rating behavior.

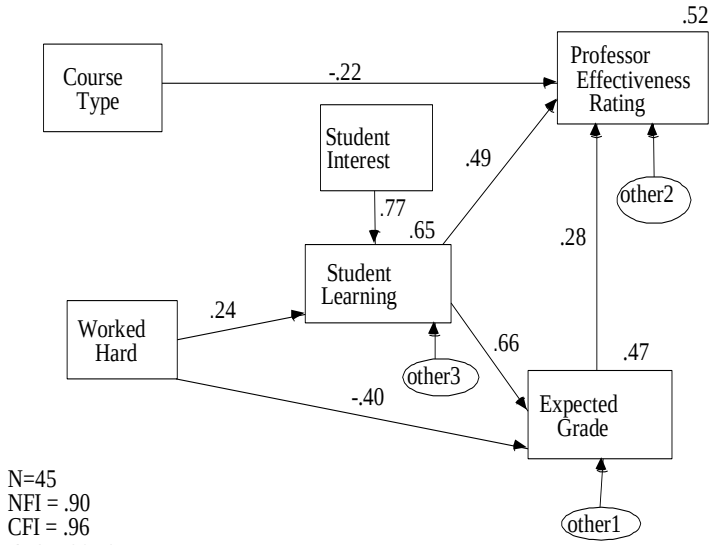
Figure 4 Integrated Model

Model A



N=45
NFI = .87
CFI = .92
AIC = 44.07
Chi-square = 16.07
df = 7
Chi-square/df = 2.30
Paths not significant:
Exp Grd --> Pro Eff
Prof Eff--> Stu Lrn
Crs Typ--> Stu Lrn
All other paths significant at alpha < .05

Model B

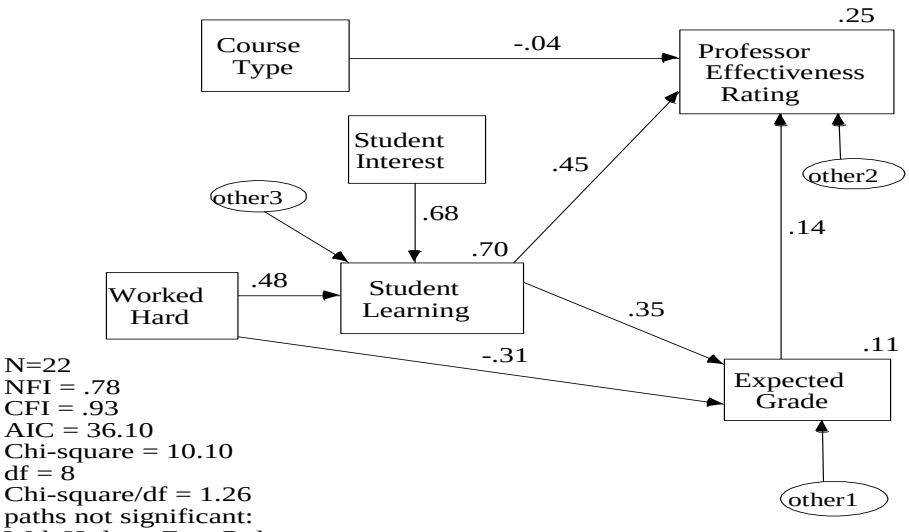


In order to assess the contribution of attribution theory in explaining the relationships found in the integrated Model B in Figure 4, the sample was split by average expected grade into a high-expected grade group and low-expected grade group. The high-expected grade model is reported in Figure 5 as Model A (CFI = .93, AIC = 36.10, R^2 = .25). The low-expected grade model is presented in Figure 6 as model A (CFI = .92, AIC = 38.36, R^2 = .64).

Inspection of the CRs for Model A in Figure 5 reveals that four paths are not significant (see Figure 5, model A). Following the previous iterative procedure used in the integrated model, these paths were dropped from the model and the final result is presented as Model B. These adjustments resulted in dropping Course Type and Expected Grade as explanatory variables. The model fits the data extremely well (CFI = .98, AIC = 17.65), all paths are significant at the .05 level, and the R^2 is .23.

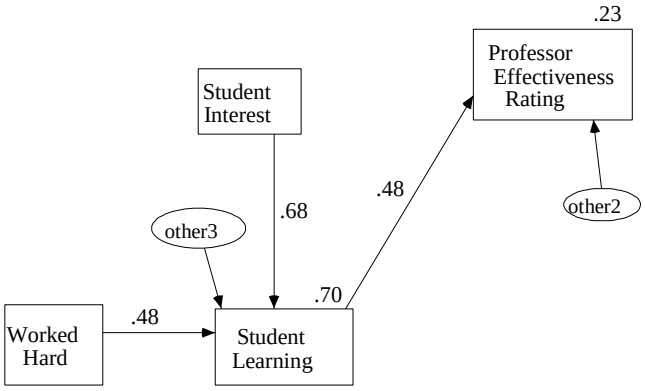
Figure 5
High Expected Grade Group

Model A



N=22
NFI = .78
CFI = .93
AIC = 36.10
Chi-square = 10.10
df = 8
Chi-square/df = 1.26
paths not significant:
Wrk Hrd --> Exp Grd
Stu Lrn --> Exp Grd
Crs Typ --> Pro Eff
Exp Grd --> Pro Eff

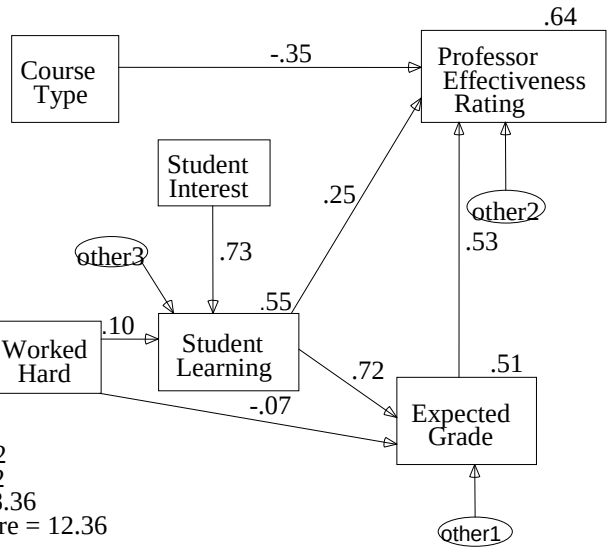
Model B



N=22
NFI = .90
CFI = .98
AIC = 17.65
Chi-square = 3.65
df = 3
Chi-square/df = 1.22
all path significant at < .05

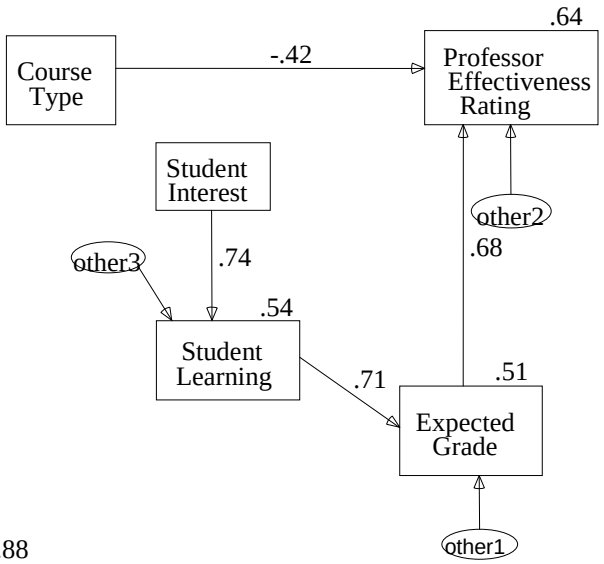
Figure 6
Low Expected Grade Group

Model A



N=23
NFI = .82
CFI = .92
AIC = 38.36
Chi-square = 12.36
df = 8
Chi-square/df = 1.55
not significant:
Wrk Hrd -->Stu Lrn
Wrk Hrd --> Exp Grd
Stu Lrn --> Pro Eff

Model B



N=23
NFI = .85
CFI = .93
AIC = 27.88
Chi-square = 9.88
df = 6
Chi-square/df = 1.65
All paths significant at < .05

Inspection of the CRs for Model A in Figure 6 reveals that three paths are not significant (see Figure 6, Model A). Following the same iterative procedure, these paths were dropped from the model and the final result is presented as model B. These adjustments resulted in dropping Worked Hard as an explanatory variable. The model fits the data very well (CFI = .93, AIC = 27.88), all paths are significant at the .05 level, and the R^2 is .64

The sharp difference in R^2s (.23 versus .64) clearly indicates the effect of attribution in student rating behavior. In the high-expected grade group, the students "attribute" their high grades to their own ability, which explains why the grade variable dropped out. Despite the fact that students learned a lot (student learning R^2 = .70), this learning is not attributed to the professor (professor effectiveness R^2 = .23). Students perceive they learned a lot because they were interested and worked hard.

For the low-expected grade group, the R^2 for the Professor effectiveness jumps to .64 (from .23 in the high-expected grade group) suggesting that the student's poorer performance is the professor's fault. This is consistent with attribution theory, since students in this group tend to blame the instructor for their poor performance, and conclude that working harder would not materially affect their grades. Students assume what they learned was due to their interest in the subject matter rather than due to the professor's teaching ability.

Several of the grade leniency linkages that were found in the integrated model in Figure 4 disappear when the sample is split into low and high-expected grade groups, e.g. Worked Hard→Expected Grade and Expected Grade→Professor Effectiveness in the high-expected grade group. This is because splitting the sample obscures the grade variance. Thus, the absence of these linkages does not alter the support found in this study for Grade Leniency theory. In fact, the Expected Grade→Professor Effectiveness linkage in the low-expected grade model suggests a grade-stringency effect. That is, students that receive lower grades rate the professor lower.

The B models in Figures 5 and 6 strongly support attribution theory. Students tend to credit themselves more and the professor less for higher learning in high-grade situations (R^2 = .23 for professor effectiveness), while blaming themselves less and the professor more for lower learning in low-grade situations (R^2 = .64 professor effectiveness). Further, the fact that the crediting/blaming occurs primarily through student learning reveals significant limitations to construct validity. Additionally, in the high-expected grade model student learning is not related to expected grade, which would be expected according to construct validity. Thus, the Attribution models in Figure 5 and 6 do not support construct validity. As in the integrated model, the high and low-expected grade models support the indirect effects of student motivation on professor effectiveness ratings. Additionally, in the low-expected grade group course type had a direct effect on professor effectiveness ratings with a relatively large standardized regression weight (|-.42|), while the path completely disappears from the high-expected grade model. That is, in low-expected grade situations, students rate professors of required courses significantly lower than professors of elective courses. It appears that students earning low grades are more influenced by course type than are students earning high grades, in terms of how they rate the professor.

Conclusions

Our findings provide strong support for the integration of grade-leniency/stringency, motivation and construct validity theories, suggesting that these theories are complementary rather than mutually exclusive (as in previous research). Student-rating behavior is a complex phenomenon. No single construct can explain it, as illustrated by the results of testing each theory independently. Additionally, this study provides strong support for attribution theory. A theory of student-rating behavior needs to integrate all four of the theoretical perspectives evaluated in this study. Integration is also necessary because of the interactions and indirect effects among the theoretical premises. For example, the attribution analysis revealed significant limitations in construct validity. That is, the variables examined in this study explained 64 percent of the variance in ratings of professor effectiveness in the low-expected grade group, and only 23 percent in the high-expected grade group. This finding clearly suggests the presence of attribution phenomenon and the presence of grade-stringency. If students earn lower grades they blame the professor and rate the professor lower, whereas if they earn higher grades they take most of the credit for themselves. Thus, strict-grading professors are rated lower than easy-grading professors. Additionally, the grade-leniency relationships were present even after the effects of motivation and construct

validity had been accounted for (see model B in Figure 4).

Although students rated professors in relation to their level of learning, learning was largely driven by course-specific interest and more generally by whether the course was an elective or a requirement. Fifty-four percent (low-expected grade group) to 70 percent (high-expected grade group) of student learning was explained primarily by student motivation and to a lesser extent by how hard students worked. Professors of elective courses were rated higher than professors of required courses.

This study found significant bias in student ratings of faculty, suggesting the need to re-consider how student ratings are used to evaluate faculty teaching effectiveness. All too often these ratings are taken at face value by evaluators of faculty, resulting in inaccurate interpretations, assessments, and comparisons. In order to more accurately evaluate professor effectiveness, administrators and faculty need to control for, or at least acknowledge, the impact of attribution, grade-leniency/stringency bias, and the effects of student motivation.

Discussion

Integration of the four theories of student rating behavior, using SEM, is a major departure from previous research that typically treated the various theoretical perspectives as mutually exclusive. This study offers a new approach and, hopefully, provides encouragement for a unified theory of student-rating behavior. The high R^2s reported here have not been reported before because so many studies (1) used correlations and occasionally regression, (2) examined a limited number of theories and theoretical premises, or (3) used within-class data, which sometimes obscures the predicted relationships because of extraneous individual or personal bias and variance. More studies using SEM are needed because of its ability to model direct and indirect cause and effect relationships. Regression is inadequate for analyzing student ratings of faculty because of the complexity of the direct and indirect relationships.

The issue of sample size is always a critical factor in any empirical study. Because of the nature of SEM analysis, this issue is particularly complex. The measurement model requires one set of guidelines and the "validation" (parameter estimates, model fit indices, etc) of the structural model requires others. Often, "rule of thumb" rather than specific criteria are use to determine acceptable sample sizes. For this study, it should be noted that we used the between-class unit of measure recommended in the literature, which helps reduce extraneous variability in the data. Further, only one of the variables (Professor Effectiveness Rating) in this study is a composite variable. All of the other variables are directly observed. Because of this design, far fewer parameters need to be estimated when using SEM. MacCallum, Roznowski, and Necowitz (1992) provide a rule of thumb suggesting five observations per parameter. Many of the models tested in this paper meet that rule and only the integrated model A (14 parameters) and model B (13 parameters) do not. In spite of the small sample size, high levels of parameter significance were obtained and strong fit indices recorded. Furthermore, R^2s recorded in this study are much higher than in earlier studies. Finally, this analysis is based on the use of the variance-covariance matrix rather than the correlation matrix. This approach is deemed more appropriate for validating causal relationships (Hair, Anderson, Tatham & Black, 1996). Thus, it can be argued that the consistency and strength of our results strongly support our findings, despite the small sample size.

Clearly our sample size of 45 between-class observations is small and larger sample sizes are needed to support and refine the results reported in this study. As a comparison, the entire within-class data set (935 observations) was used to provide estimates for the variance-covariance matrix. These values were used to repeat the analysis for the integrated model shown in figure 4, Model B. The resulting analysis yielded almost identical results. The same variables and paths appeared in both (between-class versus within-class) analyses. The path coefficients had the same signs, though their magnitudes were uniformly smaller in the with-in data analysis, reflecting larger variability. Likewise, the R^2s for the within-class were smaller (R^2 between-class = .52, R^2 within-class = .35) which also reflects the increased variability of the data. Thus, the large within-class sample results consistently supported our between-class results, with the between-class analysis (as used in this study) being the more rigorous approach.

This field would benefit from similar studies in other university settings. Also, additional work is needed on operationalizing student-rating constructs, such as general student motivation. Using SEM and approaching the analysis of student ratings of faculty from an integrated perspective should improve our understanding of student rating behavior. This research stream has matured enough that it is time to move away from correlation analysis to more robust statistical techniques and a more complex inclusive theory of student rating behavior. Researchers also need to faithfully report complete descriptive statistics and effect sizes, which has been lacking in previous studies. Such improvements should facilitate theory convergence.

Hopefully, the findings of this study, in concert with subsequent refinements and extensions, will contribute to the further development of an integrated theory of student ratings of faculty that can explain the systematic patterns and biases found in student rating behavior. The results of our research suggest that this theory will need to (1) integrate the constructs of attribution, grade-leniency/stringency, motivation, and construct validity, (2) acknowledge attribution and grade-leniency/stringency bias, and (3) acknowledge the direct and indirect effects of motivation on student behavior and ratings of professor effectiveness.

The controversy surrounding the extent of bias in student ratings needs to be re-examined. While many researchers acknowledge that bias exists, its magnitude has generally been considered too small to warrant concern. Proposed adjustments to ratings have generally been dismissed. Given the significance and magnitude of the bias found in this study, future research should focus on management and control of the bias in student ratings.

References

1. Aleamoni, L.M. 1981. Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage.
2. Anderson, J.C. & Gerbing, D.W. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3): 411-423.
3. Aronson, E. & Linder, D.E. 1965. Gain and loss of esteem as determinants of interpersonal attractiveness. *Journal of Experimental Social Psychology*, 1: 156-171.
4. Bentler, P.M. & Bonnett, D.G. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88: 588-606.
5. Bentler, P.M. & Chou, C. 1987. Practical issues in structural modeling. *Sociological Methods and Research*, 16: 78-117.
6. Bollen, K.A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
7. Braskamp, L.A. & Ory, J.C. 1994. *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.
8. Bridgeman, W.J. 1986. Student evaluations viewed as a group process factor. *Journal of Psychology*, 120: 183-190.
9. Cashin, W.E. 1988. Student ratings of teaching: A summary of the research. *Idea Paper No. 20.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
10. Cashin, W.E. 1995. Student Ratings of Teaching: The Research Revisited. *IDEA Paper No.32*. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
11. Cashin, W.E. & Downey, R. 1992. Using Global Student Rating Items for Summative Evaluation. *Journal of Educational Psychology*, *84*: 563-572.
12. Centra, J.A. 1993. *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.
13. Chacko, T.I. 1983. Student ratings of instruction: A function of grading standards. *Education Research Quarterly*, 8(2): 14-25.
14. Chapman, J.W. & Lawes, M.M. 1984. Consistency of causal attributions for expected and actual examination outcome: A study of the expectancy confirmation and egotism models. *British Journal of Educational Psychology*, 54: 177-188.
15. Cohen, P.A. 1981. Student ratings of instruction and student achievement. A meta-analysis of multisection validity studies. *Review of Educational Research*, 51: 281-309.
16. d'Apollonia, S. & Abrami, P.C. 1997. Navigating student ratings of instruction. *American Psychologist*,

52(11): 1198-1208.
17. Davis, M.H. & Stephan, W.G. 1980. Attributions for exam performance. *Journal of Applied Social Psychology,* 10: 235-248.
18. Feldman, K.A. 1976. Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4: 69-111.
19. Feldman, K.A. 1978. Course characteristics and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education,* 9: 199-242.
20. Feldman, K.A. 1989. The association between student ratings of specific instructional dimensions and student achievement: Refining an extending the synthesis of data from multisection validity studies. *Research in Higher Education,* 30(6): 583-645.
21. Gigliotti, R.J. 1987. Expectations, observations, and violations: Comparing their effects on course ratings. *Research in Higher Education*, 26: 401-415.
22. Gigliotti, R.J., & Buchtel, F.S. 1990. Attritional bias and course evaluations. *Journal of Educational Psychology,* 82: 341-351.
23. Gigliotti, R.J., & Seacrest, S.E. 1988. Academic success expectancy: The interplay of gender, situation, and meaning. *Research in Higher Education,* 29: 281-297.
24. Greenwald, A.G. & Gillmore, G.M. 1997a. Grading leniency is a removable contaminant of student ratings. *American Psychologist,* 52(11): 1209-1217.
25. Greenwald, A.G. & Gillmore, G.M. 1997b. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology,* 89(4): 743-752.
26. Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. 1998: *Multivariate Data Analysis*, 5th ed. New Jersey, Prentice Hall, 603-604.
27. Haladyna, T. & Hess, R.K. 1994. The detection and correction of bias in student ratings of instruction. *Research in Higher Education*, 35(6): 669-687
28. Hatfield, L. & Kohn, J.W. 2003. Attribution Theory Reveals Grade-Leniency/Stringency Effects In Student Ratings Of Faculty, *Academy of Educational Leadership Journal*, Vol. 7, No.2.: 1-14.
29. Holmes, D.S. 1972. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology,* 63(2): 130-133.
30. Howard, G.S. & Maxwell, S.E. 1980. Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, 72(6): 810-820.
31. Howard, G.S. & Maxwell, S.E. 1982. Do grades contaminate student evaluations of instruction? *Research in Higher Education*, 16: 175-188.
32. Howard, G.S., Conway, C.G. & Maxwell, S.E. 1985. Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology,* 77(2): 187-196.
33. Hoyt, D.P. 1973. Measurement of instructional effectiveness. *Research in Higher Education*, 1: 367-378.
34. Kennedy, W.R. 1975. Grades expected and grades received: Their relationship to students' evaluations of faculty performance. *Journal of Educational Psychology,* 67: 109-115.
35. Kline, R.B. 1998. *Principles and Practices of Structural Equation Modeling*. New York: Gilford Press.
36. Kohn, J.W. & Hatfield, L. 2001. Student Ratings of Faculty and Motivational Bias—A Structural Equation Approach, *Academy of Educational Leadership Journal*, 5(1):65-74.
37. Marsh, H.W. 1980. The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal,* 17: 219-237.
38. Marsh, H.W. 1984. Students' Evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology,* 76(5): 707-754.
39. Marsh, H.W. 1986. Self-serving effect (bias?) in academic attributions: Its relation to academic achievement and self-concept. *Journal of Educational Psychology*, 78:190-200.
40. Marsh, H.W. 1991. Multidimensional Students' Evaluations of Teaching Effectiveness: A test of Alternative higher-Order Structures. *Journal of Educational Psychology,* 83: 285-296.
41. Marsh, H.W. & Duncan, M. 1992. Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.) *Higher education: Handbook of theory and research*, 8: 143-233. New York: Agaton.
42. Marsh, H.W. & Hocevar, D. 1985. Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin*, 97(3):

562-582.
43. MacCallum, R.C., Roznowski, M. & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111(3), 490-504.
44. Marsh, H.W. & Roche, L.A. 1997. Making students' evaluations of teaching effectiveness effective. *American Psychologist*, 52(11): 1187-1197.
45. McHugh, M.C., Fisher, J.E. & Frieze, I.H. 1982. Effect of situational factors on the self-attributions of females and males. *Sex Roles*, 8:389-396.
46. McKeachie, W.J. 1997. Student ratings – The validity of use. *American Psychologist*, 52(11): 1218-1225.
47. Miller, D.C. 1991 *Handbook of Research Design and Social Measurement*. Newbury Park, California: Sage Publications, Inc.
48. Owie, I. 1985. Incongruence between expected and obtained grades and students' ratings of the instructor. *Journal of Instructional Psychology*, 12:196-199.
49. Powell, R.W. 1977. Grades, learning, and student evaluation of instruction. *Research in Higher Education*, 7: 193-205.
50. Ross, M. & Fletcher, G.J.O. 1985. Attribution and social perception. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (vol. 2, pp. 73-122). New York: Random House.
51. Simon, J.G. & Feather, N.T. 1973. Causal attribution for success and failure at university examinations. *Journal of Educational Psychology*, 64: 46-56.
52. Stumpf, S.A. & Freedman, R.D. 1979. Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology*, 71: 293-302.
53. Weiner, B. 1979. Theory of motivation for some classroom experiences. *Journal of Educational Psychology*, 71: 764-775.

Notes