Eğitim ve Bilim
2011, Cilt 36, Sayı 162

Education and Science
2011, Vol. 36, No 162

Contaminating Factors in University Students' Evaluation of Instructors

Üniversite Öğrencilerinin Öğretim Elemanlarını Değerlendirmesinde Etkili Olan Faktörler

İlker KALENDER*

Bilkent University

Abstract

The present study seeks to determine the variables explaining differences between the scores of student ratings given to instructors within the context of the university through discriminant analysis. Ratings given by students were grouped into two groups based on their means and instructors were labeled as low-rated and high-rated. Predictors identified by discriminant analysis are (i) class size, (ii) credit, (iii) grade level, (iv) mean grade, and (v) number of sections. Results of the study suggested that low rated instructors are those who teach courses with smaller number of students, lower credits, higher grade levels, higher mean grades, and one section. Identification of source of differences between ratings may provide invaluable information for those who are interested in assessment of instructional effectiveness.

Keywords: Student ratings, assessment of instructional performance, discriminant analysis

Öz

Bu çalışma, üniversite ortamında öğretim görevlilerine verilen öğrenci değerlendirme puanları arasındaki farklılıkları, ayırma analizi yöntemi ile belirlemeye çalışmaktadır. Öğrenciler tarafından verilen puanlar ortalamalarına göre iki gruba ayrılmış ve öğretim görevlileri düşük-puanlı ve yüksek-puanlı olarak iki grup halinde tanımlanmıştır. Ayırma analizi ile tanımlanan kestiriciler; (i) sınıf mevcudu, (ii) dersin kredisi, (iii) sınıf düzeyi, (iv) sınıf ortalaması ve (v) aynı dersi alan grup sayısıdır. Çalışmanın sonuçları düşük-puanlı öğretim görevlilerinin düşük sınıf mevcudu olan, düşük kredili, üst sınıf düzeylerinde, sene sonu not ortalaması yüksek olan ve tek grupla sınıflarda eğitim yapanlar olduğunu ortaya çıkartmıştır. Öğrenci değerlendirmeleri arasındaki farklılıkları açıklayan faktörlerin tanımlanması, öğretimin etkinliğinin değerlendirilmesi ve iyileştirilmesi açısından konu ile ilgili olan kişiler için önemli bulgular sağlayabilir.

Anahtar Sözcükler: Öğrenci değerlendirmeleri, öğretim performansının değerlendirilmesi, ayırma analizi.

Summary

Purpose

Assessment of effectiveness of instruction has been a concern for many years. Among several alternatives, student ratings are probably one of the most widely used measures of instructional effectiveness. Identification of factors associated with student ratings can be helpful to assess results of student ratings. The influential factors on student ratings can be considered when interpreting assessment results by using statistical control techniques or any appropriate methods. The present study seeks to determine variables explaining differences between scores of student ratings of low and high rated instructors based on selected variables from the context

* İlker KALENDER, PhD, Bilkent University, Faculty of Education, kalenderi@bilkent.edu.tr

of university setting through discriminant analysis method. A sample including mean scores of ratings of 3094 university students for 214 courses offered in the years between 2006 and 2009 was used in the present study. Confirmatory factor analysis conducted through Lisrel software (Jöreskog & Sörbom, 1999) revealed that items constituted a common factor.

Results

Low-rated instructors can be characterized courses with (i) smaller number of students, (ii) smaller credits, (iii) higher grade levels, (iv) higher mean grades, and (v) with only one sections. On the other hand, (i) higher number of students in the class, (ii) courses with higher credits, (iii) courses with lower grade levels, (iv) lower mean grades, and (v) one or more than one sections are the characteristics that define high-rated instructors. Discriminant function correctly classified 84.2% of the cases correctly. Cross-validation produced a correct classification with 83.3% hit rate.

Discussion

Results of the present study showed that students in courses with higher grade levels tend to give lowers ratings to instructors. At lower grade levels, students who are new comers to university may not evaluate instructors effectively and give higher ratings even though effectiveness of instruction is lower in the classroom. Findings of the present study indicated that instructors with higher ratings are mainly from larger classes. This can be explained by pressure on students. In small classes with student-instructor interaction, instructors may ask questions students, spend more time on their works and/or assignment, etc. much more than they do in larger classes. Credit of the course can be considered as an indicator of importance given to the course, workload needed, etc by students. When students give a course such as major courses higher importance, they may study harder, gain higher motivation and this leads to better learning and therefore higher ratings (Cashin, 1995). An explanation for lower ratings for instructors who have higher mean grades can be that following the instructors' grading policy during the semester, students expect to receive a higher grade and rate instructors positively. At the end of the semester, when they receive a lower grade since instructors assigns lower letter grades opposed to expectancies of the students, students tend to feel less satisfied. This study revealed that forty percent of the instructors with higher ratings are those who teach courses with more than one section. Students attend courses with more than one section may have instructors who devote more time to increase instructional effectiveness.

Conclusion

The present study sought to reveal the predictors that discriminate between low and high ratings given by students to instructors. As evidenced by this present study, student ratings are contaminated by several factors which means that students consider other factors intentionally as well as unintentionally. Purification or refinement of the student rating scores may be considered since student ratings is widely used for several purposes in institutions such as tenure, grants, etc. and equality among instructors or other teaching staff whose ratings can get unbalanced due to factors investigated in the present study.

Introduction

Assessment of effectiveness of instruction has been a concern for many years. Results of the assessments are used to improve teaching quality, promote faculty, pay merits, and assign of instructors to courses (Ehie & Karathanos, 1994; Kulik, 2001). Among several alternatives such as interviews with students, long-term follow-up of students, classroom visits, etc., student ratings are probably one of the most widely used measures of instructional effectiveness (Chen & Hoshower, 2003; UCLA Office of Instructional Development, 2006).

Despite its wide use, use of student rating for assessment of instructional performance is a controversial issue. There is a large accumulation of body related to validity, reliability, and contaminants that correlate with student evaluations and therefore affect their validity.

Student ratings have been under severe attack from the context of validity and reliability. Higher coefficients reported by Arubayi (1987), Costin, Greenough and Menges (1971), and Marsh (1984) provided supporting evidence for reliability of student ratings. Moreover, as stated in Aleamoni's study (1999), there are large numbers of studies indicating that scores of students can produce highly consistent reliabilities. Also studies by Cashin, Downey and Sixbury (1994), and Marsh (1984) showed that student ratings have substantial reliability. As to validity issue, there are studies providing evidence for validity of ratings (Abrami, d'Apollania & Cohen, 1990; Costin, Greennough & Menges, 1971). Though the study by Rodin and Rodin (1972) found evidence against the validity, it was criticized from the context of its methodology (Centra, 1973; Frey, 1973; Gessner, 1973; Menges, 1973).

A huge body of research focuses on the relationship between student ratings and factors associated with classroom and teacher characteristics such as gender of students/instructors, class size, grading leniency, etc. Results of the studies investigating these relationships are in conflict.

Seven studies investigated in the study by Costin, Greenough and Menges (1971) reported no correlation between gender of instructor and student ratings. On the other hand, there are lots of studies presenting results in favor of one gender of instructors (Atamian & Ganguli, 1993; Basow & Silberg, 1987; Goldberg & Callahan, 1991; Kierstead, D'Agostino & Dill, 1988; Caplan, Endres, & Lueck, 1993; Tatro, 1995). Studies cited in Aleamoni and Hexner (1980) found no relationship between class size and student ratings. On the other hand, Shapiro (1990) reported a negative correlation. Grade level of the courses is another reported factor found to be influential on student ratings. Studies in Aleamoni and Hexner (1980) found no significant relationship about student ratings. On the contrary, some studies (Conran, 1991; Donaldson, Flannery & Ross-Gordon, 1993; Goldberg & Callahan, 1991) stated that grade level is a factor affecting student ratings. Grading policy of the instructors or grading leniency is the one of most discussed factors related to student ratings. There is a belief that higher the grades students receive, higher the rating instructors receive. Studies by Greenwald and Gillmore (1997a, 1997b) indicated that grading leniency is a factor affecting student ratings. Also there are studies conducted by Endo and Della-Piana (1976), Frey (1973) which stated that student ratings were partially affected by grading leniency and might show its effect on the scores given by students. The studies reported zero correlation between student ratings and grading can also be found in the literature (Goldberg & Callahan, 1991).

Although studies in the literature have contradictory findings, many of them reveal that there are factors contaminating student ratings and therefore affecting their validity, which cannot be ignored (Kulik, 2001). The relationship between several variables related to instructor (gender, grading leniency, etc.) and course (class size, grade level, etc.) and student ratings implies that students make judgments about their instructor/courses based on different criteria beside instructional effectiveness. That is, student ratings can be said to measure something unintended as well as what they were intended to. Statistical control of these variables can be considered to remove their effects from student ratings. Cashin (1995) classified the variables with respect to requirement of control based on the literature. Age and gender of instructor were classified as instructor variables and gender of student, grade level, and mean grade as student variables requiring no control. On the other hand, faculty rank, expected grade, and level of the course were classified as variables possibly requiring control. Also Greenwald and Gillmore (1997a, 1997b) suggest applying an adjustment for the contaminating effect of the variables.

Identification of factors associated with student ratings can be helpful to assess results of student ratings. The factors that were found to be influential on student ratings can be considered when interpreting assessment results by using statistical control techniques or any appropriate methods.

The present study seeks to determine variables explaining differences between scores of student ratings of low and high rated instructors based on selected variables from the context of university setting through discriminant analysis method. Identification of predictors that explain differences between student ratings may provide invaluable information related to classroom and instructor characteristics that are potentially affecting student ratings for those who are interested in assessment and improvement of the effectiveness of instruction and other related issues.

Method

Sample

A sample including mean scores of ratings of 3094 university students for 214 courses offered in the years between 2006 and 2009 was used in the present study. Grade level covers a range between 1 and 4. There are 1791, 917, 309, and 77 students in grade levels of 1, 2, 3, and 4, respectively. Mean of the class sizes of the courses is 14.41 with a standard distribution of 5.30. Distribution of mean grade of the courses has a mean and standard deviation 2.02 (out of 4.00) and 0.59, respectively. Credit of the courses is with a mean and standard deviation 3.53 and 0.78, respectively (from 2 to 5). 159 courses are with one section (74.3%) and the rest including 55 courses (25.7%) is with more than one section.

Instrument

Instructor evaluation form filled by students includes 13 items. Items are rated using 5-point Likert type scale (from 1: Strongly disagree to 5: Strongly agree) to assess instructor performance. Students were given evaluation forms at the end of each semester before final examinations of the courses was given. No information related to identity of the students were collected and instructors did not participate evaluation sessions.

Of the 13 items in the evaluation for, mean of 6 items related to instructor performance were used as an indicator of effectiveness of instruction. The items included are (i) *Clearly states course objectives and what is expected of students,* (ii) *Stimulates interest in the subject,* (iii) *Stimulates and directs in-class student participation effectively,* (iv) *Develops students' analytical, creative, critical, and independent thinking abilities,* (v) *Interacts with students on a basis of mutual respect, and* (vi) *I learned a lot in this course.* Confirmatory factor analysis conducted through Lisrel software (Jöreskog & Sörbom, 1999) revealed that these six items constituted a common factor. Therefore this factor was accepted to be representative of instructional effectiveness. Table 1 presents the goodness-of-fit indices. Although value of RMSEA seems to be greater than acceptable threshold which is 0.05, Browne and Cudeck (1993) indicated that RMSEA values less than 0.1 indicate that model fit is not poor. Therefore based on the fit indices the conceptual model proposed is accepted.

Table 1.
Goodness of fit indices for Confirmatory Factor Analysis

Index	Value
GFI	0.94
AGFI	0.89
SRMR	0.018
RMSEA	0.090
90% confidence interval for RMSEA	(0.062; 0.12)

Analysis

In the present study, mean scores of student ratings for 214 different courses was selected as the unit of analysis rather than individual scores since students may attend to more than one evaluation sessions and also students in classroom settings can influence each other.

Ratings given by students to instructors were grouped into two groups based on means value and labeled as *low-rated (LR)* and *high-rated (HR)*. Although there seems to a large difference between groups, as one of the primary sources about discriminant analysis, Klecka (1980) did not specify any rule as to difference between groups of dependent variable. The value that discriminates between low- and hig7h-rated instructors was selected 4.00 out of 5.00. Courses with one section were coded as 1 and the courses with more than one section were coded as 2 to be able to treat it as a continuous variable.

As a first step in mean differences between LR and HR instructors were investigated to determine potentially discriminating predictors. Class size (*class*), credit of the course (*credit*), grade level of the course (*grade level*), end-of-semester mean grade of the course (*mean grade*), and number of sections category (*section*) were found to have significant mean differences between LR and HR groups ($p<0.05$).

Correlations among the predictors were checked to find whether multicollinearity exists. Mean of the correlations was found to be –0.10, which means that there is no collinearity. The highest correlation is between mean grade and the grade level of the course (0.51).

Discriminant analysis was conducted with forced-entry method with prior probabilities hold equal for both groups and significance level was set to 0.05. Also cross-validation was conducted on the data set to assess the discriminative power of the discriminant function *class*, *credit*, *grade level, mean grade*, and *section* for future cases. Cross-validation was actualized using the cases only in the analyses. Classification of the cases is made by including all cases except the one being classified.

Results

In the present study, instructors were grouped into low- and high-rated groups based on the student ratings. Predictors that discriminate between instructor ratings were determined by using discriminant analysis

Eigenvalue of the canonical discriminant function was found to be 0.31 with a canonical correlation coefficient of 0.49. Square of canonical correlation coefficient, which is 0.24, gives the percent of variance in the dependent variable explained by the set of predictors. Wilk's λ=0.765 (χ^2=56.168; df=5; p=0.00) indicated that eigenvalue for the discriminant function that explains differences between LR and HR instructors is significant. Only one discriminant function was estimated for the analysis which explains 100% of the variance.

The discriminant analysis results yielded a function of predictors that maximized the difference between LR and HR instructors. Values of Wilk's λ the standardized coefficients can be used to assess whether predictors have a significant discriminating power and determine relative importance of the predictors. The lower values of Wilk's λ indicate higher discriminating power. Therefore, among the five predictors *grade level* seemed to be predictor with the most discriminating power with a standardized coefficient of 0.915. The other predictors *class, mean, credit, grade,* and *section* had similar discriminating powers as can be seen from their standardized coefficients in Table 2.

Table 2.

Standardized Coefficients and Wilk's λ Values of Predictors

	Standardized Coefficients	Wilk's λ	F	df	Sig.
class	-0.149	0.959	9,054	212	0.003
credit	-0.287	0.960	8,941	212	0.003
grade level	0.915	0.780	59,805	212	0.000
mean grade	-0.114	0.942	13,075	212	0.000
section	-0.037	0.964	7,850	212	0.004

Discriminant analysis revealed the predictors that are sources for the differences between LR and HR instructor groups. Table 3 presents mean differences of predictors between LR and HR instructor groups.

Table 3.

Test of Equality of Group Differences for Predictors

	LR	HR
	M (SD)	M (SD)
class	11.39 (5.31)	14.83 (5.16)
credit	3.09 (0.60)	3.59 (0.78)
grade level	2.77 (0.90)	1.53 (0.71)
mean grade	2.41 (0.52)	1.97 (0.56)
section	1.00 (0.00)	1.37 (0.63)

Using Table 3, LR instructors can be characterized courses with (i) smaller number of students, (ii) smaller credits, (iii) higher grade levels, (iv) higher mean grades, and (v) with only one sections. On the other hand, (i) higher number of students in the class, (ii) courses with higher credits, (iii) courses with lower grade levels, (iv) lower mean grades, and (v) one or more than one sections are the characteristics that define HR instructors.

Discriminant function correctly classified 84.2% of the cases correctly. Cross-validation produced a correct classification with 83.3% hit rate. Results of cross-validition can be used as a measure of effect size and high hit rate indicates that discriminant function obtained in the present study could effectively be used to predict future cases and classify instructors into one of the two groups based on student ratings.

Discussion

In the present study an attempt was made to identify the variables that contaminate student ratings and, in turn, affecting their validities. To this end, discriminant analysis was conducted to identify the factors explaining mean differences in the student ratings of instructors. Predictors identified by discriminant analysis are (i) class size, (ii) credit, (iii) grade level, (iv) mean grade, and (v) number of sections of the course.

Results of the analysis suggest that students tend to rate instructors who teach courses with smaller number of students, lower credits, higher grade levels, higher mean grades, and one section, negatively. The highest contribution for the discrimination of low-rated and high-rated in-

structors comes from the predictor grade level of courses. Then comes, credit of the course, mean grade, class, credit and section of the courses, respectively,

The grade level of the course seems to be the most effective predictor discriminating between two levels of instructors. Results of the present study showed that students in courses with higher grade levels tend to give lower ratings to instructors. Results seem to support the findings of the studies by Donaldson, Flannery and Ross-Gordon (1993), Conran (1991), Goldberg and Callahan (1991), and Moritsch and Suter (1988). Scores getting higher with grade level can be explained by being mature of students over years in university. At lower grade levels, students who are new comers to university may not evaluate instructors effectively and give higher ratings even though effectiveness of instruction is lower in the classroom. Similarly at courses with higher grade level, students may get frustrated with condense curriculum and heavy workload and assign lower ratings to instructors. On the other hand, there are few studies pointing out that there is a positive (Braskamp & Ory, 1994) or no relationship (studies cited in Aleamoni & Hexner, 1980) between grade level and student ratings.

Class size is another factor that was found to be influential on student ratings. According to the findings of the present study, instructors with higher ratings are mainly from larger classes. This can be explained by pressure on students. In small classes with student-instructor interaction, instructors may ask questions students, spend more time on their works and/or assignment, etc. much more than they do in larger classes. High interaction with instructor during the semester may cause students to develop anxiety and lower his/her ratings of instructors. On the other hand, in larger classes, time of interaction per student decreases and students may feel uncomfortable and reflects this on their ratings. Generally, students in smaller classes tend to give higher ratings to instructors due to better student-instructor interaction (Mateo & Fernandez, 1996). For Turkish students, there are studies stating that students tend not to be active participants in the class, they rather prefer to passive elements in classes (Kalender & Berberoglu, 2009; Yayan & Berberoglu, 2004). Although these studies were conducted at middle-level education, they provide information about the profile of Turkish students.

Credit of the course can be considered as an indicator of importance given to the course, workload needed, etc by students. When students give a course such as major courses higher importance, they may study harder, gain higher motivation and this leads to better learning and therefore higher ratings (Cashin, 1995).

The mean grade of the course is probably the most controversial issue related to student ratings. There are studies indicating positive relationship between expected and/or actual grades and student ratings (Greenwald & Gillmore, 1997a, 1997b) as well as those that indicates no relationships (Goldberg & Callahan, 1991). Although literature generally reports positive but weak correlations between expected/actual grades and student ratings, the findings of the present study indicated that instructors with low ratings are those given higher grades. The study of Sailor, Worthen, and Shin (1997) reported similar findings. An explanation for lower ratings for instructors who have higher mean grades can be as follows: Following the instructors' grading policy during the semester, students expect to receive a higher grade and rate instructors positively and at the end of the semester receive a lower grade since instructors assigns lower letter grades opposed to expectancies of the students.

Findings of the study pointed out that 40% of instructors with higher ratings are those who teaching courses with more than one section. Students attend to courses with more than one section may have instructors who devote more time to increase instructional effectiveness.

Conclusion

The present study sought to reveal the predictors that discriminate between low and high ratings given by students to instructors. The variables that were considered to be influential on

ratings given by student to instructors for evaluating instructional effectiveness were number of students, credits of the courses, grade levels of the course, mean of end-semester grades of students, and number of sections of the courses. These variables were showed that they affected students' opinions about instructional effectiveness. A discriminant function using the variables was estimated and variables contributed the function statistically significant. High hit rate for cross-validation indicated that discriminant function works well to differentiate between instructors in terms of effectiveness.

As evidenced by the findings of the present study, student ratings are contaminated by several factors. This means that students consider other factors unintentionally or not. On the other hand, Greenwald and Gillmore (1997b) stated that even though contamination of student ratings does not necessarily mean that they are unable to measure what they are intended to measure. They may work well for the intended and unintended factors together. Purification or refinement of the student rating scores may be considered since student ratings is widely used for several purposes in institutions such as tenure, grants, etc. and equality among instructors or other teaching staff whose ratings can get unbalanced due to factors investigated in the present study. Another point that is worth to underline is that use of student ratings is only one of the several ways to evaluate instructor performance. Data should be obtained from the different sources to make sound judgments related to instructors.

The predictors identified reflect low- and high-rated instructor profiles from the perspective of Turkish university students. Studies investigated cross-cultural differences among the predictors may be conducted. Also quantitative studies to make in-depth analyses about influential factors may provide invaluable information.

The results of the present study are expected to provide valuable information for those who are interested in improving evaluation processes of instructional effectiveness.

As the literature review given in the present study indicated, findings about effects of factors influential in student ratings are contradictory. Some researchers reported significant relationships for some variables, while others do not. Therefore it is recommended to conduct research on sites where findings are intended to be applied for and factors that are influential on student ratings, if any, should be determined.

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, 82,* 219-231.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998, *Journal of Personnel Evaluation in Education, 13*(2), 153-166.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*, 67-84.

Arubayi, E. A. (1987). Improvement of instruction and teacher effectiveness: are student ratings reliable and valid. *Higher Education, 16*(3), 267-278.

Atamian, R., & Ganguli, G. (1993). Teacher popularity and teaching effectiveness: Viewpoint of accounting students. *Journal of Education for Business, 68*(3), 163-169.

Basow, S. A., & Silverg, N. T. (1987). Student evaluations of college professors: are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308-314.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, &

Long, J. S. (Eds.) *Testing Structural Equation Models*. pp. 136–162. Beverly Hills, CA: Sage.

Caplan, R. E., Endres, K. L., & Lueck, T. L. (1993). The interaction effects of gender on teaching evaluations. *Journalism Educator, 48*(3), 235-248.

Cashin, W. E., Downey, R. G. , & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives. *Journal of Educational Psychology*. *86*(4), 649-657.

Cashin, W. E. (1995). *Student ratings of teaching: the research revisited*. IDEA Paper No. 32. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80 /14/d2/44.pdf

Chen, Y. & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: an assessment of student perception and motivation, *Assessment & Evaluation in Higher Education, 28*(1), 71-88.

Centra, J. A. (1973). Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology, 65,* 395-401.

Conran, P. B. (1991). High school student evaluation of student teachers: how do they compare with professionals? *Illinois School Research and Development, 27*(2), 81-92.

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

Donaldson, J. F., Flannery, D. , & Ross-Gordon, J. (1993). A triangulated study comparing adult college students' perceptions of effective teaching with those of traditional students. *Continuing Higher Education Review*, *57*(3), 147-165.

Ehie, I. C. & Karathanos, D. (1994). Business faculty performance evaluation based on the new aacsb accreditation standards. *Journal of Education for Business*. *69*, 257-262.

Endo, G. T., Della-Piana, G. (1976). A validation study of course evaluation ratings. *Improving College and University Teaching*, 24(2), 84-86.

Frey, P. W. (1973). Student ratings of teaching: validity of several rating factors. *Science*, *182*, 83-85.

Gessner, P. K. (1973). Evaluation of instruction. *Science*, 180, 566-569.

Goldberg, G., & Callahan, J. (1991). Objectivity of student evaluations of instructors. *Journal of Education for Business, 66*(6), 377-378.

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

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,* 743-751.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8. 30*. Chicago: Scientific Software International.

Kalender, I., & Berberoglu, G. (2009). An assessment of factors related to science achievement of turkish students. *International Journal of Science Education, 31*(10), 1379 - 1394.

Klecka, William R. (1980). *Discriminant analysis. Quantitative Applications in the Social Sciences Series, No. 19*. Thousand Oaks, CA: Sage Publications.

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

Kulik, J. A. (2001). Student ratings: validity, utility, and controversy. *New Directions for Instructional Research, 27*(5), 9-25.

Lin, W. Y. (1992). Is class size a bias to student ratings of university faculty? a review. *Chinese University of Education Journal, 20*(1), 49-53.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

Mateo, M. A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement, 56*(5), 771- 778.

Menges, R. J. (1973). The new reporters: Students rate instruction. In C. R. Pace (ed.). *ew directions in higher education: evaluating learning and teaching*. San Francisco: Jossey-Bass.

Moritsch, B. G., & Suter, W. N. (1988). Correlates of halo error in teacher evaluation. *Educational Research Quarterly, 12*(3), 29-34.

Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science, 177*, 1164-1166.

Shapiro, G. E. (1990). Effect of instructor and class characteristics on students' class evaluations. *Research in Higher Education*. *31*(2), 135-148.

Sailor, P., Worthen, B., & Shin, E. H. (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. *Assessment & Evaluation in Higher Education, 22*(3), 261-269.

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research and Development in Education, 28*(3), 169-173.

UCLA Office of Instructional Development. (2006). *Guide to evaluation of instruction*. Retrieved from http://www.oid.ucla.edu/publications/evalofinstruction/index.html

Yayan, B., & Berberoglu, G. (2004). A re-analysis of the TIMSS 1999 mathematics assessment data of the Turkish students. *Studies in Educational Evaluation, 30*, 87–104.