Analysis of Bias in Student Evaluations of Faculty at an All Female Arab University in the Middle East

John Morgan and Thomas Davies

Zayed University

Learning and Teaching in Higher Education: Gulf Perspectives, Vol. 3, No. 2 (June 2006)

Abstract:

This paper reports results of analyses made at an all-female Gulf Arab university measuring the nature and extent of biases in students' evaluation of faculty. Comparisons are made with research reporting the nature of similar relationships in North America. Two issues are investigated: 1) What variables (if any) bias faculty evaluation results at an all-female Arab university? 2) Are biasing variables different in nature or magnitude to those reported at North America universities? Using the population of 13,300 faculty evaluation records collected over two school years at Zayed University, correlations of faculty evaluation results to nine potentially biasing factors are made. Results show biases to faculty evaluation results do exist. However, biases are small, and strikingly similar in nature to those reported at North American universities.

Introduction

This research investigates the nature and magnitude of grade and other biases to faculty evaluation results at Zayed University located in the United Arab Emirates. Zayed University is a national university having an all-female Gulf Arab student body. The university provides a unique and as yet uninvestigated cultural context from which to measure relationships between faculty evaluation results and a set of potentially biasing factors. Though Zayed University is a 100% female Emirati student body, it employs a Western faculty and provides students with a western style education in the English language. This unusual combination of characteristics may result in biases to the faculty evaluation process different from those found elsewhere.

A large and robust body of research exists in North America investigating issues related to biases and other shortcomings related to the faculty evaluation process. Understanding existing research has been critical to our own research design and our interpretation of results at Zayed University. North American research addresses many aspects of the faculty evaluation process including issues of construct validity, measurement reliabilities, and descriptions of the empirically observed correlations between faculty evaluation results and inappropriate biasing factors. More that three decades of research in this area have resulted in hundreds of published articles and dozens of books on this topic. What our study provides is a cross-cultural comparison of North American research results to findings at

an all-female Middle Eastern university. We have considered three broad areas of North American research in our design of this research and have compared our findings to each area.

Studies on construct validity and reliability of student ratings

A large group of North American research studies are directed at the construct validity and reliability of the student ratings process. Dozens of studies have argued both for and against the construct validity of faculty evaluation instruments in general. One intuitive, well articulated, and widely accepted view has been student ratings are valid by definition if they reflect instructional effectiveness (Abrami, Cohen, and d'Apollonia, 1988). Unfortunately, researchers rarely agree on the precise meaning of instructional effectiveness. Widely differing descriptions of instructional effectiveness have been suggested. Some place relatively more importance on mastery of knowledge and course content, others on the development of higher order intellectual skills, and others on inspiration and/or motivation of students towards continued scholarship. The meaning(s) of instructional effectiveness unavoidably must include some value judgments on the part of the person suggesting definitions (McKeachie, 1997). Scriven (1981) points out there are multiple ways to be an effective teacher. Implied in his argument is no single measure of instructional effectiveness captures all the many aspects of teaching effectiveness and therefore will be inherently flawed.

Notwithstanding unresolved ambiguities regarding definition and meaning of instructional effectiveness, the decades of the 1970s and 1980s were ones in which experts reached general agreement that results of faculty evaluations do in fact provide useful and valid insights into teacher effectiveness (Greenwald, 1997). These developments occurred primarily due to two types of evidence developed during the 1980s. One type of evidence was from large multi-section studies in which many sections of a single course were taught by multiple instructors. More than forty studies of this type were completed by the end of the 1980s (Abrami et. al., 1988). In the best of these studies students' abilities were balanced a priori using pre-testing or random assignment. Common examinations taken by all sections were used as the measure of student achievement under different instructors. This in turn permitted correlations to be made between student achievement under different instructors and student ratings of instructor effectiveness. In a large meta-analysis of these studies in 1988, Abrami et al. reported the average correlation between student achievement and instructor ratings was .40. An earlier somewhat smaller meta-analysis in 1981 by Cohen concluded overall correlation between student ratings of instructors and student achievement was at .43. Based on the totality of these studies, experts have generally concluded evidence favors the proposition that a moderate degree of correlation between student achievement and student ratings of instructor effectiveness does exist.

A second type of study has also led to wider acceptance of the validity of student ratings of instructor effectiveness. These studies report on the convergent and discriminant validities of student ratings with other factors believed to be positively or negatively associated with instructor effectiveness (Greenwald, 1997). A series of convergent validity studies showed positive correlations between student

ratings of instructor effectiveness and other indicators of instructor effectiveness including peer ratings, self-ratings, and expert judge's ratings (Harrison, Ryan, & Moore, 1996; Koon & Murray,1995; Abrami, d'Apollonia, & Cohen, 1990). Ory, Braskamp, & Pieper (1980) found positive correlations from .81 to .93 between student ratings of faculty effectiveness measured by three separate methods , objective questions, written comments, and group interviews. Fourteen studies cited by Aleamoni and Hexner (1980) showed student ratings had moderately high positive correlations with colleagues' ratings, expert judges' ratings, graduating seniors' ratings, and alumni ratings.

Discriminant validity studies also have presented evidence showing perceived biases to faculty evaluation results are not large. Factors such as expected grade, actual grade, course difficulty, and gender were all found to have small or nonexistent relationships with faculty evaluation results (Howard, Conway & Maxwell,1985; Marsh, 1982; Freidman, Stumpf, & Aguanno, 1979). Additional details of these findings are presented more fully in the next section.

Finally, research from the decade of the 1980s showed most commercially developed measurement instruments for rating of instructor effectiveness had excellent internal consistencies and high internal reliabilities in the .80 to .95 range (Arubayi, 1987; Marsh,1984). Ratings of individual instructors had high consistencies across students within a course, and high consistencies over time in longitudinal studies (Hativa, 1996; Palchic, 1988; Marsh, 1984).

Altogether, the growing body of research from North America suggested student ratings of instructor effectiveness contained some valid information even if imperfectly measured and imperfectly defined. The conclusion of a partial validity of result has become conventional wisdom over time and the research developed.

Studies on faculty perceptions about student ratings

A second type of research study from North America has been concerned with widely held perceptions about the student ratings process by faculty and students. Perhaps not surprisingly, student evaluation of instructor effectiveness has not been well received by faculty as credible or valid. Sojka, Gupta and Deeter-Schmelz (2002) reported significant differences in faculty and student perceptions regarding the means and outcomes of the student ratings process. Faculty strongly perceived students rewarded easier, more entertaining instructors. Students disagreed. Faculty and students also disagreed about whether students actually completed evaluations in a serious and careful manner. Faculty widely perceived students gave evaluations little thought before completing them. Students on the other hand perceived evaluations were taken seriously and were conscientiously completed by most students.

Aleamoni (1999) identified sixteen of the most common faculty beliefs about the student ratings process. These include perceptions that students are not serious when completing evaluations due to immaturity or capriciousness, student ratings are more reflective of popularity than of effective teaching, students are not capable of judging what is needed until they have been away for several years, student ratings are inherently unreliable and invalid measures of instructional effectiveness, grades expected by

students are strongly correlated with student ratings results, student evaluation results are skewed by class size, by gender of the instructor, by time of day a course is taught, and by whether courses are elective or required for general education or a major. Additionally, faculty indicated a belief that evaluation results are skewed by the level of the course (freshman, sophomore, junior, or senior), and by the rank of the instructor teaching the course.

In a more recent study, Baldwin and Blattner (2003) reported results of a faculty survey at a large state university in the United States. Their survey results indicated faculty believe the difficulty of the course work, initial student motivation to be in the course, pre-course interest in course material, leniency in grading, number of students in the class, and gender all unfairly bias student evaluation outcomes.

An unpublished faculty survey conducted at Zayed University in which one of the authors of this paper was involved from the 2002-2003 school year showed Zayed University faculty believed student evaluation results were unfairly biased. These results can be summarized as 79% of faculty believed results were unfairly influenced by grade, 65% believed results were biased by the content of the subject being taught, 51% by faculty gender, 50% by number of students enrolled, 48% by time of day the course was taught, and 42% by the type of course offering (i.e. general education or major requirement).

Studies showing systematic bias to the faculty evaluation process

Perhaps the largest body of published research on the faculty evaluation investigates whether or not student ratings are systematically biased by one or more inappropriate factors, and if so, the degree to which these biases exist. Relationships most often investigated and reported include connections between evaluation results and grade (expected or actual), subject matter, gender of the instructor and/or the student, class size, and time of day courses are taught.

Grade expectation

Much controversy continues to swirl around the relationship between student rating of faculty effectiveness and the actual or expected course grade. While survey research consistently shows faculty **believe** a significant correlation exists between these two factors, empirical research has been ambiguous in showing this to be the case.

Studies attempting to assess the degree of correlation between student evaluation results and expected or received course grade are mixed. Aleamoni (1999) summarized studies of this type by noting twenty-four studies show no correlation between student grade and student evaluation results; another thirty-seven studies show small positive correlations. Of the thirty-seven studies showing small positive correlations, the median correlation was only 0.14 with average variance explained by student grades less than 2%.

Multi-section research results does consistently show positive correlations between higher overall faculty evaluations by sections having higher objective overall test scores based on common tests across sections and after a priori balancing of students based on ability and interest has occurred. These

findings are used by some to argue small positive correlations sometimes found between faculty evaluation results and student grades are not from inappropriate bias but from greater learning in teacher effective environments which naturally leads to higher grades. Howard et al. (1985) argue it is logical and correct there is some positive relationship between student achievement on objective tests and student grades. In multi-section research, positive relationships between ratings of instructor effectiveness and student achievement on objective tests has been consistently established. In view of this, Howard et. al. conclude it is natural to expect that small positive correlations (appropriate correlations) between student grades and student ratings of instructor effectiveness will be observed. McKeachie (1979) using the same reasoning argues small correlations between student grades and faculty evaluation results are in fact evidence of ratings validity rather than a sign of inappropriate bias.

In short, the large body of research from North America is mixed with respect to the nature and extent of observed relationships between student grades (expected or received) and faculty evaluation results. Approximately 40% of studies find no correlation. Approximately 60% of studies show small correlations averaging only 0.14. Additionally, a convincing argument can be made based on multi-section research that small positive correlations between faculty evaluation results and student grades is justified from the objective learning differences that are known to exist across these same groups.

Subject taught

To date no research has been able to successfully establish the existence of systematic correlations between faculty evaluation results and type of course (major versus non-major). Evidence suggests an absence of this relationship (Divoky & Rothermel, 1988; Aleamoni & Thomas, 1980). Several studies do show upper-division students rate instructors slightly better than first and second year students (Conran, 1991; Moritsch & Suter, 1988). Additional studies show small differences exist across disciplines with students in humanities and social science courses rating instructors slightly better than students in quantitative and physical science courses. (Andrew, Gauthier, & Jelmberg, 1993; Goodwin & Stevens, 1993; Cashin, 1990).

Gender

A notably large number of studies have been conducted on gender and its relationship to faculty evaluation results. Gender of faculty, gender of students, and interactions between the two have all been investigated. Results of these studies are contradictory and highly confusing. The preponderance of studies show no systematic differences between male and female students in their evaluations of faculty as a main effect. They also show no differences in evaluations received by male and female faculty as a main effect (Basow, 1995; Amin, 1994; Feldman, 1993; Aleamoni & Thomas, 1980).

Summers, Anderson, Hines, Gelder, and Dean (1996) reported contrary results showing both male and female students rate female instructors slightly lower than male instructors. Tatro (1995) and Kierstead, d'Agostino, and Dill (1988) and found exactly the opposite. Their results showed female

instructors were rated higher than male instructors by both male and female students. Baslow (1995) reported no main effects for either student or faculty gender, but did find a significant interaction showing female students rate female faculty somewhat higher. No conclusions can be easily drawn from such conflicting data except the effects, when found are small.

Class size

Intuition might suggest smaller classes would be evaluated more favorably than larger classes because smaller classes permit more faculty-student interaction and more personal attention than would otherwise be possible. Research results are nevertheless mixed about this. Aleamoni and Hexner (1980) cite seven studies finding no relationship between class size and student evaluation results, and also cite eight additional studies showing a weak but nevertheless systematic relationship between class size and student evaluation results in the direction that instructors in smaller classes are rated better. More recent studies by Lin (1992) and Shapiro (1990) have also confirmed weak correlations between smaller classes and higher faculty evaluation results.

Time of day

Aleamoni (1999) reports there has been little work done on the relationship between time of day a class is taught and student evaluation results. Two studies of this type, Feldman, (1978) and Yongkittikul, Gillmore, and Brandenburg (1974) found no connection between the two variables.

Empirical relationships observed at Zayed University

Relationships between faculty evaluation results and nine potentially biasing factors were investigated at an all-female Arab university (Zayed University) located in the Middle East. Results from the population of 13,300 individual faculty evaluation records completed by students over three semesters during 2004 and 2005 are compared to the nine potentially biasing factors. Operationally, faculty evaluation results are defined as the simple average of two summative questions found on the faculty evaluation instrument used by Zayed University. Both questions concern instructor effectiveness. The two questions are: 1) *The overall effectiveness of the instructor is*:, and; 2) *I would tell other students that the instructor is*:. Students responded to these items using a one to five scale with five meaning *excellent*, four meaning *good*, three meaning *average*, two meaning *weak*, and one meaning *poor*. The mean faculty evaluation result in the 13,330 records examined was a rather high 4.36 and the standard deviation was .78. Nine potentially biasing variables were compared with students' ratings of instructor effectiveness. The nine variables were:

Grade received.

Subject matter (coded as primarily quantitative or primarily non-quantitative).

Course level (coded as a 100, 200, 300, or 400 level course)

Gender of faculty member (all students were female)

Number of students enrolled in the section.

Time of day (coded as middle of the day, early, or late).

Baccalaureate course offering type (coded general education or major).

Faculty member length of experience at Zayed University (coded as first semester or beyond first semester).

Student level at Zayed University (credits already completed).

Using correlation coefficients for formal analysis

Correlation has been the statistic most widely used in research literature for evaluating the nature and strength of observed relationships between student ratings of faculty effectiveness with other studied variables such as grade received, subject matter, course level, gender of teacher, and class size. In order to facilitate the comparison of our research findings with other studies, we have chosen to use the same statistical measure (i.e. correlation coefficient) for evaluating the relationship between variables.

A correlation coefficient (r), is descriptive statistic measuring the degree to which there is systematic covariation between two things. *Correlation coefficients* are stated as a number between negative one and one and measure the degree to which two items of vary together in a systematic way. If there is a perfect positive relationship between two items (i.e. as one item doubles, so does the other) the correlation coefficient is positive one. If there is a perfect negative correlation between two items (i.e. as one item doubles and the other is halved) the correlation coefficient is negative one.

Unfortunately, relationships observed between items inferred from very small samples may not closely approximate the true relationship in the population from which they have been drawn. Statisticians have long understood the likelihood of observed relationships from samples accurately representing true population relationships depends in part upon the sample size.

A simple illustration helps explain this principle. Assume a large bowl contains 100,000 balls of which 10,000 are white and 90,000 are red. If just two balls are drawn from the bowl (a small sample of two) and both are red, one might erroneously think the bowl contains only red balls. Because the sample size was only two balls, the conclusion is not reliable in the statistical sense. On the other hand, if 5,000 balls are drawn from the bowl (sample size of 5,000) and approximately 10% are white and the other 90% are red, the conclusion that approximately 10% and 90% represent true population proportions is much more reliable in the statistical sense. Mathematical statisticians have developed formulas that precisely measure the probabilities associated with samples of a given size (in a given context) accurately representing underlying population characteristics. Interestingly, with **very** large samples it becomes possible to identify even tiny systematic differences between two variables, and with a high degree of reliability.

Correlation coefficients developed from a sample of a given size have a related 'significance statistic' that expresses the mathematical probability that observed covariations are indeed systematic

and representative of the underlying population, and not from random chance. In most scientific studies, a significance statistic of less than 5% (sometimes less than 1%) is considered acceptable for reaching a research conclusion. It is easier to understand 'significance statistic' using its complement. For example, the more direct interpretation of a 5% significance statistic is that one can be 95% confident the related correlation coefficient measures a real systematic relationship, not one based on chance alone. Significance, in this sense, refers to statistical probability and should not be confused with practical significance or magnitude of a relationship. It should be noted when using very large samples (as in this study), even tiny correlations (very weak relationships) can be identified with a high degree of statistical confidence. In this study, only those relationships between variables having a significance statistic of less than 1% (i.e. 99 percent confidence level) are reported as statistically significant.

One final advantage of using the correlation coefficient as the measure of relationship is that it provides a convenient way to express the strength of relationships between variables (i.e. variance explained). When a correlation coefficient, r, is .40, *variance explained* by the relationship is, by definition, 'r-squared' [.40 X .40 = .16]. R-squared provides a direct expression of explained percentage of covariance between two variables which in this example is only 16%. The other 84% of covariance is **not** explained by any systematic relationship between the two variables and thus results from causes unknown. When reporting our research findings, we report the correlation coefficient, its related significance statistic, and the percentage of covariance explained with this relationship (r-squared).

Grade received

The correlation between faculty evaluation results and grade received was statistically significant at Zayed University. SPSS table details are included in the appendix. Students receiving higher grades rated faculty slightly more effective than otherwise. The correlation between faculty evaluation results and student grade received at Zayed University was .17 (sig. < .01 meaning more than 99% confident the relationship is not chance related to small sample size). Variance explained (r-squared) by the association was only 3%. The other 97% of covariance between student evaluation results and grade received is from other unknown sources. Average scores received by faculty from students with an "A" grade was 4.52, with a "B" grade was 4.37, with a "C" grade was 4.15, with a "D" grade was 4.08, and with an "F" grade was 3.98. The level of correlation between faculty evaluation results and grade received appears remarkably similar in size to relationships reported by North American researchers in the studies conducted at North American universities (Aleamoni ,1999).

Subject matter (quantitative or qualitative)

For purposes of determining biases to faculty evaluation resulting from the qualitative versus quantitative nature of courses, we first coded courses as either primarily qualitative or quantitative. This classification was necessarily somewhat subjective. All natural sciences, mathematics, statistics, research methods, computer programming, accounting, finance, and economic courses were coded as

quantitative. All others courses in the humanities and most social sciences courses were coded as qualitative. The correlation between faculty evaluation results and course type (qualitative or quantitative) was statistically significant at Zayed University but very very small. SPSS table details are included in the appendix. The correlation coefficient between faculty evaluation results and course type was a mere .03 (significance statistic < .01). Variance explained by the association was less than one tenth of one percent. Overall, qualitative courses were evaluated slightly higher than quantitative courses. The mean faculty evaluation result in qualitative courses was 4.37 and in quantitative courses was 4.30.

Gender of faculty member

Gender of faculty member was determined to be a statistically significant correlate to faculty evaluation results at Zayed University though once again very very small. SPSS table details are included in the appendix. Correlation coefficient was .05 (significance statistic < .01) explaining less than one tenth of one percent of the total variance. Overall, male faculty were evaluated slightly higher than female faculty. The mean faculty evaluation result for male faculty was 4.39 and for female faculty was 4.30.

Number of students enrolled

At Zayed University even the largest classes are small in number compared to what is often the case at large public universities. Data showed the largest class size over the last three semesters at Zayed University had only 34 students. Perhaps not surprisingly, class size (which was uniformly small) was not a statistically significant correlate of faculty evaluation results at Zayed University.

Time of day

A preliminary look at data suggested to the authors instructors in early classes and late classes were similarly evaluated lower than middle of the day classes. For purposes of our study, start times were categorized as either *middle of the day* or *not the middle of the day*. Middle of the day classes were defined as those classes starting between 9:00 a.m. and 2:00 p.m. Not middle of the day classes were defined as classes starting before 9:00 a.m. or after 2:00 p.m.

Correlation between faculty evaluation results and start times was statistically significant but again exceedingly small with a correlation coefficient of only 0.07 (significance statistic < .01). Variance explained by this covariation was less than one percent of the total. The other ninety-nine percent was due to other factors. SPSS table details are included in the appendix. Overall, instructors in middle of the day classes were evaluated slightly higher than instructors in early and late classes. The mean faculty evaluation result for middle of the day classes was 4.40 and for early or late day start was 4.29.

Course offering type (general education or major requirement)

Courses were categorized as either fulfilling a general education requirement or as being taken to fulfill a major requirement. The data showed class type (general education or major requirement) was not a statistically significant correlate of faculty evaluation results at Zayed University.

Course level code (100, 200, 300, 400)

Research at North American universities has shown a tendency by third and fourth year students to evaluate faculty slightly higher than first and second year students. This tendency was not found at Zayed University. The correlation between faculty evaluation results and course level code (100, 200, 300, 400) was not statistically significant.

Faculty experience at Zayed University

Because student background and the teaching environment at Zayed University is considerably different from what many faculty have experienced in their home nations, it may be the case that beginning faculty have some initial difficulties adjusting to learning styles at Zayed University. To test this hypothesis, faculty were divided into two groups: first semester faculty and faculty beyond the first semester. The correlation between faculty evaluation results and faculty experience level at Zayed University was statistically significant. SPSS table details are included in the appendix. Faculty in their first semester received slightly lower evaluation results than those in the second semester and beyond. The effect however was again very small. The correlation coefficient for the relationship between faculty evaluation results and faculty experience (first semester or beyond) was only 0.04 (significance statistic < .01). Variance explained by the association was just two tenths of one percent, hardly worth considering. The mean faculty evaluation result for first semester faculty was 4.28 and for faculty beyond the first semester was 4.37.

Prior credits completed

Data on the number of prior credits completed by students and faculty evaluation results were not significantly correlated at Zayed University. No systematic relationship between number of semesters completed by students and evaluation results was found.

Multivariate model (stepwise model)

In addition to the nine separate univariate comparisons reported above, a multivariate model was created for evaluating the best multivariate relating the nine potentially biasing factors with faculty evaluation results. This model was created using stepwise regression whereby variables are entered one at a time into the model and are later removed if they do not significantly contribute to the multiple regression result. Variables with the strongest univariate relationship were entered first. The final

stepwise model included only five of the nine variables: grade received, start time, gender, experience of faculty member, and course type (quantitative or qualitative). The final stepwise model is summarized as (SPSS table details are included in the appendix):

Evaluation Results = Constant + ß(Grade Points) + ß(Start Time)
+ ß(Faculty Experience) + ß(Faculty Gender)
+ ß(Course Type)

Evaluation Results = 3.365 + .155(Grade Pts) + .098(Start Time)
+ .098(F. Experience) + .084(Faculty Gender)
+ 062(Course Type)

While the model itself was statistically significant (significance statistic < .01), the magnitude of the overall multivariate association was still small. Variance explained by the five variables together in the multivariate model was only 3.8% which is only slightly higher than the 3% explained by the single variable, *grade received*, alone.

Discussion and conclusions

Of the nine potentially biasing variables investigated at Zayed University, five variables had statistically significant but very small univariate relationships with faculty evaluation results. Four others had no systematic relationships. Correlated with faculty evaluation results were grade received, subject matter (coded as quantitative or qualitative), faculty gender, start time, and faculty experience. Uncorrelated with faculty evaluation results were class size, type of course offering (general education or major requirement), course level (100, 200, 300, 400), and semesters completed at Zayed University by the evaluating student.

The magnitude of univariate and multivariate relationships between the five statistically significant variables and faculty evaluation results were all small, and similar to those reported at North American educational institutions. A summary of univariate relationships found at Zayed University listed in order of size is listed in Table 1 below:

TABLE 1 GOES HERE

Grade received has the single strongest individual correlation with faculty evaluation results. However at 0.17, the correlation only predicts three percent of the total variance to faculty evaluation results. The other 97% of the relationship results from things other than grade received. Also as has already been noted above North American researchers have concluded, within the context of multi-section research studies, a logical reason does exist for positive correlations between student grades

and faculty evaluation results. Under conditions of more effective instruction, student learning is greater, and thus grades are higher. In short, one might logically expect a small positive relationship between student achievement and student grades, unrelated to pandering on the part of the instructor. Many see small positive correlations between student grades and student ratings of instructors as evidence of the construct validity of faculty evaluation results rather than evidence of their bias. In any case, the strength of the association between grade received and faculty evaluation results at Zayed University is not large and is similar in strength to those reported above by North American researchers.

Virtually none of the other statistically significant variables in Table 1 account for even one-half of one percent of the total variance in faculty evaluation results. Whatever the specific nature of the individual relationship between each variable and evaluation results, the magnitudes are so small they make virtually no difference.

Finally, as noted earlier the best multivariate model developed collectively from all five significant variables accounts for only 3.9% of the total variance in faculty evaluation results. Accordingly, the authors conclude, while systematic biases to the faculty evaluation process at Zayed University can be identified in a very large sample (n = 13,300), they are individually and collectively small, relatively unimportant overall, and are remarkably consistent with the small levels of bias reported at North American educational institutions over the last several decades. Based on these small correlations between nine potentially biasing variables and student evaluations of faculty effectiveness at Zayed University, an all-female Arab university, we conclude as have experts in other contexts, there is little evidence of major inappropriate biases to faculty evaluation results.

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 know. *Journal of Educational Psychology,* 82(2), 219-231.

Abrami, P.C., Cohen, P.A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research,* 58, 151-179.

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.

Aleamoni, L. M., & Thomas, G. S. (1980). Differential relationships of student, instructor, and course relationships to general and specific items on a course evaluation questionnaire. *Teaching of Psychology*, 7(4), 233-235.

Amin, M. E. (1994). Gender as a discriminating factor in the evaluation of teaching. *Assessment and Evaluation in Higher Education*, 19(2), 135-143.

Andrew, M. D., Gauthier, S. A., & Jelmberg, J.R. (1993). Comparing student perceptions of instruction in teacher education and education courses. *Journal of Personnel Evaluation in Education*, 6(4), 359-366.

Arubayi, E. A. (1987). Improvement of instruction and teacher effectiveness: are student ratings reliable and valid? *Higher Education*, 16(3), 26-28.

Baldwin, T., & Blattner, N. (2003). Guarding against potential bias in student evaluations. *College Teaching*, 51(1), 27-33.

Baslow, S. A. (1995). Student evaluations of college professors. *Journal of Educational Psychology*, 87(4), 656-665.

Cashin, W. E. (1990). Students do rate academic fields differently. *New Directions for Teaching and Learning (Student Ratings of Instruction: Issues for Improving Practice)*, 43, 113-121.

Centra, J. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, 14, 17-24.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies. *Review of Educational Research*, 51, 281-309.

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

Divoky, J. J. & Rothermel, M. A. (1988). Student perceptions of the relative importance of dimensions of teaching performance across type of class. *Educational Research Quarterly*, 12(3), 40-45.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: what we know and what we don't. *Research in Higher Education*, 9, 199-241.

Feldman, K. A. (1993). College students' views of male and female college teachers: part II, evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34(2), 151-211.

Freedman, R. D., Stumpf, S. A., & Aguanno, J. C. (1979). Validity of the course-faculty instrument (CFI): intrinsic and extrinsic variables. *Educational and Psychological Measurement*, 39, 153-158.

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

Harrison, P.D., Ryan, J. M., & Moore, P. S. (1996). College students' self-insight and common implicit theories in ratings of teaching effectiveness. *Journal of Educational Psychology*, 88(4), 775-782.

Hativa, N. (1996). University instructors' ratings profiles: stability over time, and disciplinary differences. *Research in Higher Education*, 37(3), 341-365.

Howard, G. S., Conway, C. G. & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, 77, 187-196.

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.

Koon, J., & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education*, 66(1), 61-81.

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. (1982). Validity of students' evaluations of college teaching: a multitrait-multimethod analysis. *Journal of Educational Psychology*, 74, 264-279.

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

McKeachie, W. J. (1979). Student ratings of faculty: a reprise. *Academe*, 65, 384-397.

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

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

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology*, 72(2), 181-185.

Palchik, N. S. (1988). Student assessment of teaching effectiveness in a multi-instructor course for multidisciplinary health professional students. *Evaluation and Health Professions*, 11(1), 55-73.

Scriven, M. (1981). Summative teacher evaluation. *Handbook of Teacher Evaluation*, Beverly Hills, CA: Sage, 244-271.

Shapiro, E. G. (1990). Effects of instructor and class characteristics on students' class evaluation. *Research in Higher Education*, 3(2), 135-148. Sojka, J., Gupta, A. K., Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: a study of similarities and differences. *College Teaching*, 50(2), 44-51.

Summers, M. A., Anderson, J. L., Hines, A. R., Gelder, B. C., & Dean, R. S. (1996). The camera adds more than pounds: gender differences in course satisfaction for campus and distance learning students. *Journal of Research and Development in Education*, 29(4), 212-219.

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

Yongkittikul, C., Gillmore, G. M., & Brandenburg, D. C. (1974). Does the time of course meeting affect course ratings by students? *Research Report No. 346,* Urbana: University of Illinois, Office of Instructional Resources, Measurement, and Research Division.

About the authors

John Morgan received his Ph.D. in accountancy from the University of Nebraska-Lincoln. He is a certified public accountant and a certified management accountant and has worked as a university professor in North America and the Middle East for the past twenty-six years, the last five years being at Zayed University.

Thomas Davies is Head of Institutional Research at Zayed University.

Appendix of SPSS tables

Linear Regression Summary (SELE Result with Student Grade Received)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.172(a)	.030	.030	.76866

a Predictors: (Constant), Grade Points

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.875	.025		154.331	.000
	Grade Points	.157	.008	.172	20.074	.000

a Dependent Variable: ***SELE Result***

Mean *SELE Result* by *Student Grade Received*

	N	Mean	Std. Deviation	Std. Error
A	3672	4.5234	.67194	.01109
B+	2787	4.3967	.75456	.01429
B	3169	4.3593	.75665	.01344
C+	1496	4.2246	.81398	.02104
C	1274	4.1527	.90367	.02532
D+	237	4.1624	.94890	.06164
D	373	4.0818	.90586	.04690
F	200	3.9825	.96337	.06812
Total	13208	4.3605	.78027	.00679

==========

Linear Regression Summary (*SELE Result* with *Course Type)*

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.032(a)	.001	.001	.77991

a Predictors: (Constant), Course Type

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.438	.022		198.443	.000
	Course Type	-.067	.018	-.032	-3.666	.000

a Dependent Variable: ***SELE Result***

Mean *SELE Results* by *Course Type*

	N	Mean	Std. Deviation	Std. Error
Qualitative	11176	4.3704	.77051	.00729
Quantitative	2154	4.3032	.82701	.01782
Total	13330	4.3596	.78028	.00676

Linear Regression Summary (*SELE Results* with *Faculty Gender*)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.055(a)	.003	.003	.77912

a Predictors: (Constant), Faculty Gender

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.298	.012		364.466	.000
	Faculty Gender	.091	.014	.055	6.360	.000

a Dependent Variable: ***SELE Result***

Mean *SELE Result* by *Faculty Gender*

	N	Mean	Std. Deviation	Std. Error
Female	4365	4.2981	.81788	.01238
Male	8965	4.3895	.75954	.00802
Total	13330	4.3596	.78028	.00676

Linear Regression (*SELE Result* with *Start Time*)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.066(a)	.004	.004	.77863

a Predictors: (Constant), Start time

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.179	.025		169.113	.000
	Start Time	.108	.014	.066	7.586	.000

a Dependent Variable: ***SELE Result***

Mean *SELE Result* by *Start Time*

	N	Mean	Std. Deviation	Std. Error
Early Day or Late Day Start	4472	4.2876	.80796	.01208
Middle of the Day Start	8858	4.3959	.76339	.00811
Total	13330	4.3596	.78028	.00676

==========

Linear Regression (*SELE Result* with *Faculty Experience*)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.035(a)	.001	.001	.77909

a Predictors: (Constant), experience

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.247	.029		147.516	.000
	Faculty experience	.065	.016	.035	4.075	.000

a Dependent Variable: ***SELE Result***

Mean *SELE Result* by *Faculty Experience*

	N	Mean	Std. Deviation	Std. Error
First year faculty	3158	4.3114	.82830	.01474
Beyond first year faculty	10124	4.3761	.76309	.00758
Total	13282	4.3608	.77954	.00676

Multiple Regression Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Stepwise (final model)	.196	.039	.038	.76454

Predictors: (Constant), Grade Points, Start Time, Faculty Gender , Faculty Experience, Course Type

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	3.406	.057		59.255	.000
Grade Points	.155	.008	.170	19.812	.000
Start Time	.102	.014	.062	7.207	.000
Faculty Gender	.082	.014	.049	5.708	.000
Faculty Experience	.075	.016	.041	4.773	.000
Course Type	.064	.018	.030	3.500	.000

Dependent Variable: *SELE Result*

TABLE 1 Summary of Univariate Relationships

VARIABLE **NAME**	**LEVEL OF CORRELATION WITH FACULTY EVALUATION RESULTS**	**VARIANCE EXPLAINED BY VARIABLE**
1. Grade received	0.172	0.030
2. Start Time	0.066	0.004
3. Faculty Gender	0.055	0.003
4. Faculty Experience	0.040	0.002
5. Subject Matter	0.032	0.001