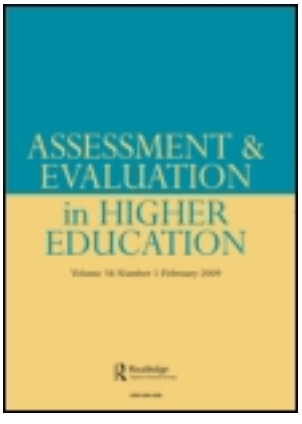

Assessment & Evaluation in Higher Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caeh20

Stability and correlates of student evaluations of teaching at a Chinese university

Guo-Hai Chen [a] & David Watkins [b]

[a] School of Management, Guangdong University of Foreign Studies, Guangdong, China

[b] Faculty of Education, The University of Hong Kong, Hong Kong

Available online: 11 Oct 2010

To cite this article: Guo-Hai Chen & David Watkins (2010): Stability and correlates of student evaluations of teaching at a Chinese university, Assessment & Evaluation in Higher Education, 35:6, 675-685

To link to this article: http://dx.doi.org/10.1080/02602930902977715

Assessment & Evaluation in Higher Education
Vol. 35, No. 6, October 2010, 675–685

Routledge
Taylor & Francis Group

Stability and correlates of student evaluations of teaching at a Chinese university

Guo-Hai Chen[a]* and David Watkins[b]

[a]*School of Management, Guangdong University of Foreign Studies, Guangdong, China;*
[b]*Faculty of Education, The University of Hong Kong, Hong Kong*

This paper examines the stability and validity of a student evaluations of teaching (SET) instrument used by the administration at a university in the PR China. The SET scores for two semesters of courses taught by 435 teachers were collected. Total 388 teachers (170 males and 218 females) were also invited to fill out the 60-item NEO Five-Factor Inventory together with a demographic information questionnaire. The SET responses were found to have very high internal consistency and confirmatory factor analysis supported a one-factor solution. The SET re-test correlations were .62 for both the teachers who taught the same course ($n = 234$) and those who taught a different course in the second semester ($n = 201$). Linguistics teachers received higher SET scores than either social science or humanities or science and technology teachers. Student ratings were significantly related to Neuroticism and Extraversion. Regression results showed that the Big-Five personality traits as a group explained only 2.6% of the total variance of student ratings and academic discipline explained 12.7% of the total variance of student ratings. Overall the stability and validity of SET was supported and future uses of SET scores in the PR China are discussed.

Keywords: Chinese university; student evaluations of teaching; personality

Introduction

Student ratings of teaching effectiveness (student evaluations of teaching, i.e. SET scores) were introduced in several major US universities in the 1920s (d'Apollonia and Abrami 1997) and had become widely used by the 1980s as an easy and objective means for universities in many areas of the world to assess teaching quality (Pounder 2007).

As discussed in this paper, the use of such ratings is controversial despite a considerable amount of supporting findings based on Western research. Student ratings have also recently been introduced in many universities of the PR China (PRC), but as yet there are fewer empirical data related to such usage. The aim of this paper is to provide evidence of the stability and teacher and institutional correlates of the SET scores as used at one Chinese university.

Reliability and validity of SET scores

Although theories of teaching and learning supported by factor analytic studies typically support a multidimensional model of teaching (cf. Marsh and Dunkin 1992, for

*Corresponding author. Email: mypeer2002@hotmail.com

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930902977715
http://www.informaworld.com

a review), in practice short undimensional scales are usually used as in this research. A large volume of research has examined the reliability and validity of student ratings of teaching effectiveness, and a number of reviews have reported that student ratings are reasonably reliable and valid and relatively free from bias (e.g. Marsh 1987; Marsh and Dunkin 1992; Centra 1993; d'Apollonia and Abrami 1997; Wachtel 1998).

Reliability

Reliability of SET scores is usually assessed in terms of internal consistency and stability over time. As Marsh and Dunkin (1992) concluded, 'internal consistency among responses to items designed to measure the same component of effective teaching is consistently high'. Correlating SET scores by the same lecturer in two different courses or in the same course over time can not only provide evidence of the generalisability of such scores but also help sort out the influence of the course and teacher effects, crucial to the interpretation of such scores. Murray, Rushton, and Paunonen (1990) analysed ratings of 46 psychology instructors in up to six courses over several years. They found high correlations averaging .86 for the same teacher teaching the same course over several years. This figure dropped only marginally when different courses were considered. Murray, Rushton, and Paunonen (1990) concluded, as had Marsh and Overall (1981) that the teacher rather than the course is the main influence on SET scores.

Validity

According to Feldman's (1986) review of this research, student ratings have been shown to have the following average correlations with external indices of teaching effectiveness: .29 with instructor ratings, .39 with administrator ratings, .55 with colleague ratings and .40 with student achievement. These correlations provide support for the construct validity of student ratings.

Wachtel (1998) also concluded that after nearly seven decades of research on the use of student ratings, the majority of researchers believe that student ratings are a reliable, valid and worthwhile means of evaluating teaching. However, many other researchers have expressed reservations about their use, particularly for professional and career decisions. Some have even opposed student ratings outright (e.g. Chandler 1978; Vasta and Sarmiento 1979; Dowell and Neal 1982; Small, Hollenbe, and Haley 1982; Miller 1984; Koblitz 1990; Rutland 1990; Zoller 1992; Goldman 1993). In particular, there is abundant anecdotal evidence of faculty hostility and cynicism towards the use of student ratings (Franklin and Theall 1989; Pounder 2007). The main criticisms claim that students will give higher SET scores to teachers who require little homework and grades to teachers who teach small or optional classes.

Factors that affect student ratings

Much research has been conducted on the relationship between student ratings of teaching effectiveness and a variety of situational variables, teachers' demographic variables and personal characteristics. Factors examined in this research are briefly reviewed below.

Teacher gender

Findings on the relationship between gender and student ratings of teaching effectiveness are quite mixed. Some studies have contended that student ratings are biased against female instructors, and that female teachers have to behave in stereotypically feminine ways in order to avoid receiving lower ratings than male teachers (e.g. Bennett 1982; Basow and Silberg 1987; Kierstead, D'Agostino, and Dill 1988; Koblitz 1990; Rutland 1990). Basow and Silberg (1987) found that male students gave female professors significantly poorer ratings than they gave male professors. On the other hand, other studies have reported little gender bias (e.g. Elmore and Lapointe 1974; Stratham, Richardson, and Cook 1991; Feldman 1993; Marsh et al. 1997; Liaw and Goh 2003).

Academic rank

Findings of research on this topic are again mixed. Several studies found that full, associate and assistant professors were rated more highly by students in comparison with teaching assistants (Centra and Creech 1976; Brandenburg, Slinde, and Batista 1977; Marsh and Dunkin 1992). However, Feldman (1983) concluded that the majority of studies showed no significant correlation between academic rank and student ratings. Feldman also found that the majority of studies yielded no significant correlation between age/experience of the teacher and student ratings.

Academic discipline

Research has also been carried out into how student ratings may be affected by the academic discipline of the subject matter, and the findings indicate that teachers in humanities and social sciences usually receive higher ratings than those in engineering and sciences (Feldman 1978). More recently, Cashin (1990) analysed the course average ratings of different fields of study for two large sets of student ratings in the USA and divided the ratings into three groups: high, medium and low. He found that consistent with Feldman's (1978) findings, the arts and humanities were likely to fall into the 'high' group and English language, literature, history, social sciences and health most often fell into the 'medium' group, whereas business, economics, computer science, mathematics, physical sciences and engineering most often fell into the 'low' group.

Class size

A significant relationship between class size and student ratings has been found in previous studies both in Western and Eastern countries. For example, Centra and Creech (1976) found a clear curvilinear relationship between class size and student ratings. They found that classes in the 35–100 range had the lowest ratings, whereas larger and smaller classes received higher ratings. Their findings were supported by Marsh, Overall and Kesler (1979) and Feldman (1984). Marsh et al. (1997) reported that class size was modestly correlated with ratings of good teachers. Good teachers with larger class sizes were rated somewhat lower in terms of Group Interaction and Individual Rapport while for poor teachers the correlations with class size were smaller. Kwan (1999) found significant differences among class size groups in a Hong

Kong university: smaller classes tending to receive higher ratings than larger classes, which was supported by Liaw and Goh's (2003) findings in a Malaysian university. Chen (2000) reported a significantly negative correlation ($r = -.25$, $p < .05$) between class size and student ratings with a sample of 93 courses in a Southern China university. It can be generally concluded that there exists a weak and yet significantly negative correlation between class size and student ratings (Marsh 1987). Although student ratings can be affected by a number of factors including teachers' demographic characteristics and situational characteristics, Kwan (1999) concluded that academic discipline and class size had particularly large effects on student ratings in terms of effect size.

Teacher personality

As teaching is partially a social or an interpersonal process, it is reasonable to hypothesise that teachers' personality traits may correlate significantly with student ratings of teaching effectiveness. Traits such as extraversion, leadership, objectivity, lack of anxiety, dominance, supportiveness, potency, intellectual competence and warmth have been found to have significant relationships with student ratings (Murray 1975; Sherman and Blackburn 1975; Tomasco 1980; Bennett 1982; Rushton, Murray, and Paunonen 1983; Murray, Rushton, and Paunonen 1990; Erdle, Murray, and Rushton 1985; Radmacher and Martin 2001). For example, Murray (1975) found that colleague ratings of instructor extraversion, leadership, objectivity and lack of anxiety accounted for 67% of the total variance in student ratings of teaching. Moreover, by using colleagues as judges, Murray, Rushton, and Paunonen (1990) found that teaching effectiveness in a range of courses could be predicted with considerable accuracy from colleague ratings of personality.

Expressiveness is one of the personality traits of teachers, which has been found to have a significant relationship with student ratings of teaching effectiveness (e.g. Basow and Silberg 1987; Basow 1998). A major criticism of such ratings is known as the Dr Fox Effect (Marsh and Ware 1982). Critics argue that enthusiastic lecturers can 'seduce' students into giving favourable evaluations, even though the lectures may be devoid of meaningful content. Ware and Williams (1975, 1976), for example, found that higher achievement was associated with high content coverage and high expressiveness and students gave higher ratings to expressive lecturers. They concluded that student ratings of highly expressive instructors do not reflect two important dimensions of teaching effectiveness, namely, substantiveness of instruction and the degree of student achievement. Their findings have been supported by several other studies (Basow and Distenfeld 1985; Basow 1990; William and Ceci 1997). However, by reanalysing data from two studies by Ware and Williams (1975, 1976), Marsh and Ware (1982) showed that for students in an incentive condition the Dr Fox effect was not supported as instructors' expressiveness only affected ratings of their enthusiasm. However, when students were not given an incentive to learn, instructor expressiveness had a greater impact on each of the student rating factors and examination performance than did content coverage.

Big-Five personality traits

Although no studies have been found so far which explore the relationship between teachers' Big-Five personality traits and student ratings, it has been suggested that

students generally tend to prefer lecturers with personality characteristics similar to theirs. For example, with a sample of 136 university students, Furnham and Chamorro-Premuzic (2005) found that students' Extraversion scores were significantly and negatively correlated with their preferences for neurotic lecturers; significantly and positively correlated with preferences for open lecturers and significantly and negatively correlated with preferences for agreeable lecturers. Students' Openness scores were found to be significantly and positively correlated with preferences for open lecturers. Students' Agreeableness scores were found to be significantly and positively correlated with preferences for agreeable lecturers. There was also a modest but significant positive correlation between students' Conscientiousness and their preferences for agreeable lecturers. It can be inferred that lecturers tend to get higher scores from students with personality characteristics similar to theirs than from those with different personality characteristics.

Aims of research

Although there were several empirical studies (e.g. Marsh et al. 1997; Kwan 1999) in Hong Kong, universities in Hong Kong are quite different from those in mainland China and more Westernised in terms of lecturers' and students' experiences. Since there are a few and valid empirical studies on student ratings in PRC, we are interested in examining if the use of SET in PRC is reliable and valid and there are similar factors that affect SET scores as in the West. This research utilised actual student ratings of teaching effectiveness collected by the administration of one PRC university. The reliability and validity of the SET scores were assessed by investigating their internal consistency, reliability, factor structure, stability over time and correlates with teacher gender, age, academic degree, academic rank, class size, academic discipline and Big-Five personality traits. Such data provide the first evidence of the use of SET scores in China and can provide insights into the worth of such in a non-Western context. The study would offer us several findings in PRC and we can compare them with those of the West in this area.

We consider it is important to examine the relationship between the teachers' Big-Five personality traits, a personality model recognised and used world-wide (McCrae and Allik 2002), and student ratings of teaching effectiveness since findings on this topic may help teachers improve their teaching effectiveness by strengthening or weakening certain aspects of their personality traits. Correlates with other characteristics may also indicate in what ways the SET scores are biased or at least need to be interpreted with care.

Method

Participants and procedure

About 7560 students at one teaching and research oriented university in the PRC filled out anonymous SET forms at the end of each semester during the academic year 2003–04. Four hundred and thirty-five teachers were rated by students on each of the courses they taught. For this research each teacher was asked to select one of their courses for first semester. If the teacher retaught that course to different students in the second semester it was chosen again for the second semester otherwise they were asked to select another course they were teaching. SET data and class size for each

Table 1. Demographic data of the teacher participants ($N = 388$).

Categories	Participant no.	(%)
Gender		
Male	170	(43.8)
Female	218	(56.2)
Academic degree		
BA or equivalent	89	(22.9)
MA/PhD or equivalent	299	(77.1)
Academic rank		
Teaching assistant	111	(28.6)
Lecturer	134	(34.5)
Associate professor	111	(28.6)
Professor	32	(8.2)
Academic discipline		
Linguistics	178	(45.9)
Social sciences and humanities	168	(43.3)
Science and technology	42	(10.8)

course were collected from the Dean's Office of the university. In total 234 teachers taught the same course during the two semesters while 201 teachers taught a different course.

Three hundred and eighty-eight teachers were also invited to fill out the 60-item NEO Five-Factor Inventory (NEO-FFI; Costa and McCrae 1992) together with a demographic information questionnaire (concerning gender, age, academic degree, academic rank and academic discipline in the semester when the survey was conducted). Demographic details of the teacher participants are shown in Table 1.

Measures

Student evaluations of teaching scale

Since 2002, all the teaching staff at this university were required to be evaluated by the undergraduate students enrolled in their courses and students were asked to rate their teachers on a five-point Likert scale, from '1 = very poor performance' to '5 = excellent performance'. This scale was developed by the first author and his former colleagues and contains eight items. Typical items are 'The teacher prepares lessons well' and 'The teacher inspires students' creativity'.

NEO Five-Factor inventory (NEO-FFI)

This is a brief but comprehensive measure of the Big-Five personality dimensions (Costa and McCrae 1992), each of which is assessed by 12 items. Each item is rated on a five-point Likert scale ranging from 1 ('not true') to 5 ('very true'). The present study employed a Chinese translation of the NEO-Five-Factor inventory. The Cronbach's alpha coefficients reported in the previous studies (Zhang and Huang 2001; He 2005) were a bit low for the Openness to experience subscale, ranging from .52 to .56, while the Cronbach's alpha coefficients for the other four subscales ranged from .64 (Agreeableness) to .85 (Neuroticism): acceptable for group research purposes.

Data analysis

To evaluate the psychometric properties of the SET scores, firstly Cronbach's alpha and then stability correlates from one academic term and those of the following academic term were computed. Confirmatory factor analysis was conducted to test a one-factor solution under the SET scores. Correlations, Analysis of Variance (ANOVA), and *t*-tests were used to investigate factors related to the SET scores.

Results

Responses to the SET scale were found to have a very high internal consistency (alpha = .98) and the confirmatory factor analysis supported a one-factor solution (CFI = .98; GFI = .95; RMSEA = .07; $n = 7560$): very satisfactory fit indices according to Hair et al. (1998). The SET re-test correlations were .62 for both the teachers who taught the same course ($n = 234$) and for those who taught a different course in the second semester ($n = 201$).

There were no significant differences in the SET scores according to the gender of the teacher (for males, $M = 86.19$, $SD = 5.52$; for females, $M = 86.93$, $SD = 5.37$, $t = 1.33$, 386 *df*, $p > .05$). There was also no significant relationship between the age of the teachers and their SET scores ($r = -.023$, 388 *df*, $p > .05$). No significant differences in the SET scores based on academic degree and rank were found. One-way ANOVA of the SET scores according to the academic discipline indicated that the SET scores in the three academic discipline groups were significantly different ($F = 15.53$, 2, 385 *df*, $p < .001$). The linguistics teachers received higher SET scores ($M = 88.2$, $SD = 4.7$) than either the social science or humanities ($M = 85.5$, $SD = 5.8$, $p < .05$) or the science and technology teachers ($M = 84.3$, $SD = 5.3$, $p < .05$). Student ratings were significantly related to class size ($r = -.18$, $p < .01$).

Student ratings were significantly related to Neuroticism ($r = -.13$, $p < .01$) and Extraversion ($r = .12$, $p < .05$) but no significant correlations were found between these ratings and any of the other three personality traits ($r = -.07$, .09 and .07 for Openness to experience, Agreeableness and Conscientiousness, respectively, all $p > .05$).

The stepwise multiple regression results showed that the Big-Five personality traits as a group explained only 2.6% of the total variance of student ratings with gender, age, class size and academic discipline controlled. Neuroticism explained 1.4% of the total variance of student ratings (F (2, 387) = 18, $p < .001$, $\beta = -.12$, $t = -2.5$, $p < .05$), while the other four personality traits (namely Extraversion, Openness to experience, Agreeableness and Conscientiousness) did not significantly contribute to the prediction of student ratings when Neuroticism was accounted for. Gender, age, class size and academic discipline could explain 14.2% of the total variance of student ratings. Academic discipline explained 12.7% of the total variance of student ratings (F (4, 387) = 19, $p < .001$, $\beta = -.34$, $t = -5.4$, $p < .001$), while the other three (namely gender, age and class size) did not significantly contribute to the prediction of student ratings.

Discussion

More similarities than differences in findings of SET were found between the present study and previous empirical studies in the West. The results supported the internal consistency reliability and unidimensionality of the SET scores as utilised in this PRC university. The scores were also fairly stable over the two semesters irrespective of

whether the lecturer was teaching the same or a different course in Semester 2. Thus, the results were consistent with those of Marsh and Overall (1981) and Murray, Rushton, and Paunonen (1990) in finding that SET scores were more influenced by the lecturer rather than the course(s) he or she was teaching.

Significant differences in the SET scores based on academic discipline were found in the present study, similar to the Western literature (Feldman 1978; Cashin 1990). So when raw data of student ratings are interpreted, the university administrators are needed to be aware of the effects of academic discipline. As Kwan (1999) stated, judgement or decisions based on raw student ratings will be likely at fault unless such potential biases are taken into consideration. Comparison of the raw data of student ratings in a variety of academic fields, currently a common practice in the excellent teaching awards of the higher education institutions in mainland China, needs to be done with caution. It is suggested that we should adjust the SET scores according to the historical data and academic discipline, and distribute limited quotas of excellent teachers in different academic fields and disciplines. Although the effects of class size on student ratings has been found to be significant in studies elsewhere (e.g. Marsh 1987; Watkins and Afzulpurkar 1988) and class sizes were apparently different in the three academic disciplinary areas (namely social sciences and humanities, linguistics, and science and technology) at this PRC University, the regression results show that SET scores were more affected by academic discipline than class size.

The personality of the teacher as measured by the Big-Five personality traits also showed little relationship with the SET scores, with only Neuroticism and Extraversion providing small statistically significant correlations. Although the significant correlation between Extraversion and SET scores is in line with the previous findings (e.g. Murray 1975; Radmacher and Martin 2001), self-ratings of teacher personality traits contributed little to the prediction of SET scores. This finding is different from that of past research that colleague ratings of teacher personality traits considerably predicted the SET scores (e.g. Murray 1975; Murray, Rushton, and Paunonen 1990; Radmacher and Martin 2001). One possible explanation is that there are significant differences between self- and colleague-ratings of teacher personality on the same characteristics.

The finding that SET scores are influenced by the teacher but not to any great extent by their Big-Five personality traits, suggests that some narrower personality variables may significantly affect SET scores (e.g. credibility, Beatty and Zahn 1990). Each Big-Five factor of personality is a broad measure incorporating a number of narrower personality variables. For example, most researchers agree that Conscientiousness measures a multifaceted array of more specific traits that include thoroughness, reliability and perseverance, as well as their opposites, carelessness, negligence and unreliability (e.g. Costa and McCrae 1992). As such, it makes sense for future research to explore the relationships between more specific facets of the Big-Five factors and student ratings.

The finding that SET scores are influenced by the teacher but not to any great extent by their personality, also suggests that other than teacher personality, some aspects such as teaching styles influence SET scores (e.g. Hudak and Anderson 1984). Future PRC research can address this issue and this would also need to be cross-validated in the Western context.

Conclusion

These results indicate that SET scores with very adequate psychometric properties were being utilised by the administration of this PRC university. Moreover, as with

the US findings, it seems that the ratings are more dependent on the teacher rather than the particular course they were teaching. There was also no evidence of possible biases in the ratings according to the gender, age or academic rank of the teachers. Personality correlates of the SET scores were not strong but did suggest as expected that less neurotic and more extroverted teachers were rated more highly. There was also some evidence of disciplinary differences which would need to be considered when decisions are being made on the basis of the SET scores.

The above findings do support the validity of SET scores in PRC universities, but exactly how the scores are being used for employment-related decisions needs further research. More SET research in PRC needs to be reported and comparison of findings in this area between PRC and West needs to be done further. Moreover, a multidimensional instrument measuring different aspects of teaching along the lines that Marsh and Dunkin (1992) proposed needs to be developed and validated for the PRC context to provide feedback for improving university teaching.

Notes on contributors

Guo-Hai Chen is a full professor of management at School of Management, Guangdong University of Foreign Studies, Guangzhou, China. His research focus has been on organisational behaviour, HR management, coaching and humour from psychological and educational perspectives.

David Watkins is a full professor at Faculty of Education, The University of Hong Kong, Hong Kong. His research focus has been on educational psychology and cross-cultural research.

References

Basow, S.A. 1990. Effects of teacher expressiveness: Mediated by teacher sex-typing? *Journal of Educational Psychology* 82: 599–602.

Basow, S.A. 1998. Student evaluations: The role of gender bias and teaching styles. In *Arming Athena: Career strategies for women in academe,* ed. L.H. Collins, J.C. Chrisler, and K. Quina, 135–56. Thousand Oaks, CA: Sage.

Basow, S., and M.S. Distenfeld. 1985. Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology* 77: 45–52.

Basow, S.A., and N.T. Silberg. 1987. Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology* 79, no. 3: 308–14.

Beatty, M.J., and C.J. Zahn. 1990. Are student ratings of communication instructors due to 'easy' grading practices? An analysis of teacher credibility and student-reported performance levels. *Communication Education* 39, no. 4: 275–91.

Bennett, S.K. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology* 74: 170–9.

Brandenburg, D.C., J.A. Slinde, and E.E. Batista. 1977. Student ratings of instructor: Validity and normative interpretations. *Research in Higher Education* 7: 67–78.

Cashin, W.E. 1990. Students do rate different academic fields differently. In *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning,* Vol. 43, ed. M. Theall and J. Franklin, 113–21. San Francisco, CA: Jossey Bass.

Centra, J.A. 1993. *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco, CA: Jossey-Bass.

Centra, J.A., and F.R. Creech. 1976. *The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. Project report, 76–1.* Princeton, NJ: Educational Testing Service.

Chandler, J.A. 1978. The questionable status of student evaluations of teaching. *Teaching of Psychology* 5: 150–2.

Chen, G.H. 2000. An empirical study on some issues of teaching evaluation by university students. Paper presented at the International Conference 'New Millennium: Quality & Innovations in Higher Education', 4–5 December 2000 in Hong Kong.
Costa Jr., P.T., and R.R. McCrae. 1992. *The NEO-PI-R: Professional manual.* Odessa, FL: Psychological Assessment Resources.
d'Apollonia, S., and P.C. Abrami. 1997. Navigating student ratings of instruction. *American Psychologist* 52, 1198–208.
Dowell, D.A., and J.A. Neal. 1982. A selective review of the validity of student ratings of teaching. *Journal of Higher Education* 53: 51–62.
Elmore, P.B., and K.A. Lapointe. 1974. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology* 66, no. 3: 386–9.
Erdle, S., H.G. Murray, and J.P. Rushton. 1985. Personality, classroom behavior, and student ratings of college teaching effectivness: A path analysis. *Journal of Educational Psychology* 77: 394–407.
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: 199–242.
Feldman, K.A. 1983. Seniority and experience of college teachers as related to evaluations they receive. *Research in Higher Education* 18: 3–124.
Feldman, K.A. 1984. Class size and students' evaluation of college teachers and courses: A closer look. *Research in Higher Education* 21: 45–116.
Feldman, K.A. 1986. The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education* 24: 139–213.
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: 151–211.
Franklin, J., and M. Theall. 1989. Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction. Paper presented at the Annual Meeting of the American Education Research Association, March 27–31, in San Francisco.
Furnham, A., and T. Chamorro-Premuzic. 2005. Individual differences in students' preferences for lecturers' personalities. *Journal of Individual Differences* 26, no 2: 176–84.
Goldman, L. 1993. On the erosion of education and the eroding foundations of teacher education (or why we should not take student evaluation of faculty seriously). *Teacher Education Quarterly* 20: 57–64.
Hair, J.F., R.E. Anderson, R.L. Tatham, and W.C. Black. 1998. *Multivariate data analysis.* New York: Macmillan.
He, Y.F. 2005. The roles of thinking styles in learning and achievement among Chinese university students. PhD Diss., The University of Hong Kong.
Hudak, M.A., and D.E. Anderson. 1984. Teaching style and student ratings. *Teaching of Psychology* 11, no. 3: 177–8.
Kierstead, D., P. D'Agostino, and H. Dill. 1988. Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology* 80: 342–4.
Koblitz, N. 1990. Are student ratings unfair to women? *Newsletter of the Association for Women in Mathematics* 20: 17–9.
Kwan, K.P. 1999. How fair are student ratings in assessing the teaching performance of university teachers? *Assessment & Evaluation in Higher Education* 24, no. 2: 181–95.
Liaw, S.H., and K.L. Goh. 2003. Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management* 17, no. 1: 37–43.
Marsh, H.W. 1987. Students' evaluation of university teaching: research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.
Marsh, H.W., and M.J. Dunkin. 1992. Students' evaluation of university teaching: A multidimentional perspective. In *Higher education: Handbook of theory and research,* Vol. 8, ed. J. C. Smart, 143–234. New York: Agathon.
Marsh, H.W., K.T. Hau, C.M. Chung, and T.L.P. Siu. 1997. Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality instrument. *Journal of Educational Psychology* 89, no. 3: 568–72.

Marsh, H.W., and J.U. Overall. 1981. The relative influence of course level, course type, and instructor on students' evaluations of college teaching. *American Educational Research Journal* 18, no. 1: 103–12.
Marsh, H.W., J.U. Overall, and S.P. Kesler. 1979. Class size, students' evaluation, and instructional effectiveness. *American Educational Research Journal* 16: 57–70.
Marsh, H.W., and J.E. Ware. 1982. Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology* 74: 126–34.
McCrae, R.R., and J. Allik. 2002. *The five-factor model of personality across cultures.* New York: Kluwer Academic/Plenum.
Miller, S.N. 1984. Student ratings scales for tenure and promotion. *Improving College and University Teaching* 32: 87–90.
Murray, H.G. 1975. Predicting student ratings of college teaching from peer ratings of personality types. *Teaching of Psychology* 2: 66–70.
Murray, H.G., J.P. Rushton, and S.V. Paunonen. 1990. Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology* 82, no. 2: 250–61.
Pounder, J.S. 2007. Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education* 15, no. 2: 178–91.
Radmacher, S.A., and D.J. Martin. 2001. Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology: Interdisciplinary & Applied* 135, no. 3: 259–68.
Rushton, J.P., H.G. Murray, and S.V. Paunonen. 1983. Personality, research creativity, and teaching effectiveness in university professors. *Scientometrics* 5: 93–116.
Rutland, P. 1990. Some considerations regarding teaching evaluations. *Political Science Teacher* 3: 1–2.
Sherman, B.R., and R.T. Blackburn. 1975. Personal characteristics and teaching effectiveness of college faculty. *Journal of Educational Psychology* 67: 124–31.
Small, A.C., K.A.R. Hollenbe, and R.L. Haley. 1982. The effect of emotional state on student ratings of instructors. *Teaching of Psychology* 9: 205–8.
Stratham, A., L. Richardson, and J.A. Cook. 1991. *Gender and university teaching.* Albany: State University of New York Press.
Tomasco, A.T. 1980. Student perceptions of instructional and personality characteristics of faculty: A canonical analysis. *Teaching of Psychology* 7: 79–82.
Vasta, R., and R.F. Sarmiento. 1979. Liberal grading improve evaluations but not performance. *Journal of Educational Psychology* 71: 207–11.
Wachtel, H.K. 1998. Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education* 23, no. 2: 191–211.
Ware, J.E., and R.G. Williams. 1975. The Dr. Fox effect: A study of lecturer expressiveness and ratings of instruction. *Journal of Medical Education* 50: 149–56.
Ware, J.E., and R.G. Williams. 1976. Validity of student ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. *Journal of Educational Psychology* 68, no. 1: 48–56.
Watkins, D., and N. Afzulpurkar. 1988. Class size and student ratings of tertiary courses. *Educational and Psychological Measurement* 48, no. 2: 523–6.
Williams, W.M., and S.J. Ceci. 1997. 'How'm I doing'? Problems with student rating of instructors and courses. *Change* 29, no. 5: 12–23.
Zhang, L.F., and J.F. Huang. 2001. Thinking styles and the five-factor model of personality. *European Journal of Personality* 15: 465–76.
Zoller, U. 1992. Faculty teaching performance evaluation in higher science education: Issues and implications (a 'cross-cultural' case study). *Science Education* 76: 673–84.