Are Teacher Course Evaluations Biased Against Faculty That Teach Quantitative Methods Courses?

Kenneth D. Royal[1] & Myrah R. Stockdale[2]

[1] Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA

[2] Department of Educational Research Methodology, School of Education, University of North Carolina at Greensboro, Greensboro, NC, USA

Correspondence: Kenneth D. Royal, Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, 1060 William Moore Dr., Raleigh, North Carolina 27695, USA. E-mail: kdroyal2@ncsu.edu

Received: January 18, 2015 Accepted: February 2, 2015 Online Published: February 3, 2015

doi:10.5430/ijhe.v4n1p217 URL: http://dx.doi.org/10.5430/ijhe.v4n1p217

Abstract

The present study investigated graduate students' responses to teacher/course evaluations (TCE) to determine if students' responses were inherently biased against faculty who teach quantitative methods courses. Item response theory (IRT) and Differential Item Functioning (DIF) techniques were utilized for data analysis. Results indicate students in non-methods courses preferred the structure of quantitative courses, but tend to be more critical of quantitative instructors. Authors encourage consumers of TCE results to investigate item-level results, as opposed to summative results, when making inferences about course and instructor quality.

Keywords: Teacher evaluation, Course evaluation, Faculty, Bias, Student ratings, Measurement

1. Introduction

The digital age has come about rapidly and has unapologetically left much of the "old" workforce in its wake. In order to keep up with the technological and information boom, many have had to learn new skills in order to thrive. One of the most in-demand skills resulting from the information age has been a fluency with quantitative literacy, as quantitative skills are necessary to successfully interpret and interact with the changing world and the seemingly infinite amount of data within it.

Recognizing the need for quantitative skills in the workforce, many colleges and universities are now making quantitative reasoning and literacy a learning outcome as part of the general education requirements(Reason, Terenzini, & Domingo, 2006; Rhodes, 2010). With the thrust in education and society toward science, technology, engineering, and mathematical (STEM) fields, it is becoming increasingly common for students to be exposed to multiple quantitative courses during both their undergraduate and graduate studies. Unfortunately, many students tend to be less enthusiastic about courses that focus on quantitative content, such as statistics, measurement, quantitative research methods, and so on (Dunn, 2000). Additionally, students who have negative mathematical or statistical experiences tend to display high levels of statistical anxiety (Mji & Onwuegbuzie, 2004). This disinterest and anxiety could manifest itself into teacher and course evaluations (TCEs) and result in a significant problem for faculty that teach these types of courses.

Research has revealed that TCEs are no longer used for their original purposes of monitoring teaching quality and improving teaching (Kulik, 2001). Many, if not most, institutions routinely use TCE scores to determine hiring, tenure and promotion, and merit-based pay decisions. It stands to reason that faculty who teach quantitatively oriented courses could be disadvantaged with regard to many administrative policies if students are indeed less enthusiastic about these courses and provide biased ratings on TCEs. The purpose of this study was to systematically investigate thousands of student ratings on TCEs in a large college of education to determine if response bias was present in courses that focused on quantitative methods and approaches.

1.1 Review of Selected Literature

There have been a myriad of studies examining student evaluation ratings of instructors. Centra (2003) stated that student evaluations of instruction have been researched more than any other topic in the college instruction literature. In spite of this, only a small portion of these studies have specifically investigated the potential for rating bias across different academic fields (Centra, 1993; Feldman, 1978). Most notably is Cashin's (1990) study that examined a large sample of students using the Students Instructional Ratings questionnaire (SIR) and the Individual Development Evaluation Assessment (IDEA) scales, both of which have been utilized in many colleges and universities. Cashin, later corroborated by others in the literature (Francis, 2011; Franklin & Theall, 1992; Gilroy, 2003; Kember & Leung, 2011), demonstrated that students do indeed invite bias into their ratings, consistently rating instructors who teach quantitative-based courses (e.g., math, hard sciences) lower than instructors who teach more qualitative courses (e.g., the arts and humanities). This finding also corresponded with Feldman's original study in 1978, as well as another study by Barnes and Patterson (1988).

Some proposed explanations have been conveyed in the literature for this consistent finding. Centra (2009) discusses the reason for rating bias across academic disciplines may be that instructors who teach quantitative-based courses tend to be in fields that progress much faster and are more competitive for research funding than other "softer" and less quantitative-heavy fields and thus, these issues may subsequently affect the teaching ability of professors, due to not devoting as much time and energy into teaching the material, whether intentionally or unintentionally. There is also evidence that differences may arise in teaching methodology, given that the hard sciences/quantitative-centered courses tend to be more lectured-oriented and less collaborative with other students than the softer sciences/humanistic courses (Centra, 2003; Marsh, 2007; Neumann & Neumann, 1985).

Students who lack a solid foundation in math and quantitative reasoning are found to be more likely to develop negative attitudes and beliefs towards quantitative reasoning, mathematics, and statistics (Earp, 2007; Jordan & Haines, 2003). These same students carry an exorbitant amount of anxiety, with as much as 80% of all students experiencing "statistic anxiety" (Hanna, Shevlin, & Dempster, 2008; Onwuegbuzie & Wilson, 2003). This enduring, habitual type of anxiety has been supported by corresponding retest correlations (Macher, Paechter, Papousek, & Ruggeri, 2012; Papousek, Ruggeri, Macher, Paechter, Heene, Weiss, et al, 2012). Undergraduate students with moderate to high anxiety perform at lower levels, resulting in lower GPAs, than students with low anxiety (Chapell, Blanding, Silverstein, Takahashi, Newman, Gubi, et al., 2005). Statistics anxiety is associated with two factors that cannot be controlled by faculty: personal background (gender, ethnicity, socioeconomic status) and prior educational experiences (success/failure in primary education) (Earp, 2007; Hendel, 1980; Richardson & Woolfolk, 1980). The third factor associated with statistics anxiety, learning motivation, can be influenced by an instructor. Learning motivation is contingent upon quality of instruction, personality and attitude of instructor and difficulty of the material. Anxiety paired with fear of failure could prove to be a plausible explanation for negative skew in quantitative course TECs however, the classroom environment itself could add to the narrative.

It is realized that there are other potential factors affecting student ratings (e.g., class size, grading leniency, level of course, etc.) (Brockx, Spooren, & Mortelmans, 2011; Pritchard, & Potter, 2011). For example, Haladyna and Hess (1994) explored and confirmed bias among student ratings using the many-faceted Rasch Model (Linacre, 1987), specifically looking into five facets of rating bias, those being the faculty member, the survey items themselves, rater bias (student), gender of the rater, rater perception of course type, and whether the course was a required course or an elective course. It was found that all elements displayed some element of bias, although perceived course type did not specifically indicate differences in academic discipline. There have been a host of models also developed in the literature as well, although examining all of these in an exhaustive fashion is beyond the scope of this study (Narayanan, Sawaya, & Johnson, 2014).

Many authors state the specific need for subsequent research in evaluating bias in student ratings of quantitative courses. Marsh and Roche (1997) also states that there is not enough evidence to make any definitive conclusions on potential student rating bias against faculty who teach quantitative courses opening up an opportunity for additional studies. Ramsden (1991) states that there may be too much variability across or even within departments to produce reliable rating results, yet it has been demonstrated that using student evaluation rating scales is a credible way to assess teaching ability among academic instructors, at least for the psychometrically established student evaluation instruments (e.g., SIR) (d' Apollonia & Abrami, 1997; Kember & Leung, 2008; Marsh, 1984; Marsh & Roche). An extensive and heavily cited review by Wachtel (1998) discussed the need to further investigate the variability of ratings within departments. Recently, Kember and Leung (2011) demonstrated significant differences of hard versus soft science through structural equation modeling, still showing that even today these differences still exist. The lack

of conclusive evidence regarding student ratings between fields and the disagreement of scholars within the student rating literature prompts the need for a study using advanced methodological techniques that allow objective measurement to answer this intriguing phenomenon of student rating bias among quantitative instructors.

1.2 The Present Study

The present study investigated graduate students' responses to teacher/course evaluations (TCE) in the College of Education at a large Midwestern university. In particular, we sought to determine if graduate students' ratings of faculty who teach quantitative methods courses were significantly different than ratings for faculty teaching other types of courses. To investigate this question we utilized the Rasch Rating Scale Model, a powerful IRT technique, and Differential Item Functioning (DIF) methods.

2. Method

Perhaps the most effective way to investigate if any biases are present in TCEs is to utilize an item response theory (IRT) approach for data analysis. IRT methods are considered particularly advantageous as they are capable of producing a great deal of information about individual persons and items within any dataset. While a discussion of IRT and Rasch measurement is beyond the scope of this article, readers are encouraged to read Bond and Fox (2007) and Royal (2010) for a thorough overview of IRT and Rasch measurement in the context of survey research.

2.1 Instrumentation

The Teacher/Course Evaluation (TCE) is not unlike most end-of-the-semester course and/or instructor evaluations. The instrument used in the present study was a uniform instrument administered throughout the university. The instrument contained 19 items partitioned among three categories: A) Course items (n=8); B) Instructor items (n=6); and, C) Learning outcomes (n=5). A five-point rating scale was provided, with response options *Strongly Disagree*, *Disagree*, *Agree*, *Strongly Agree*, and *Not Applicable*. Table 1 presents the survey items appearing on the TCE.

Table 1. Survey Items

Items
Q1 – Instructor outlined in reasonable detail course material and grading procedures
Q2 – Textbook contributed to my understanding of the subject
Q3 – Assignments helped me to understand the subject
Q4 – Exams reflected what was taught in the course
Q5 – Grading was fair and consistent
Q6 – Assignments were distributed fairly throughout the semester
Q7 – Graded assignments were returned promptly
Q8 – Graded assignments included helpful comments from the instructor
Q9 – Instructor presented material in an effective manner
Q10 – Instructor had a good knowledge of the subject matter
Q11 – Instructor was available for consultation during office hours
Q12 – Instructor satisfactorily answered questions raised in class
Q13 – Instructor stimulated my interest in the subject
Q14 – Instructor encouraged student participation in class
Q15 – I learned to respect viewpoints different from my own
Q16 – Course strengthened my ability to analyze and evaluation information
Q17 – Course helped me to develop the ability to solve problems
Q18 – I gained an understanding of concepts and principles in this field
Q19 – Course stimulated me to read further into the area

2.2 Data Acquisition

At the end of each semester, students are asked to complete Teacher/Course Evaluations (TCE) for each course that they are currently enrolled. While a summary of results are publicly available, the raw data are not. Records of the raw data are maintained by the university's Office of Institutional Research (IR). Upon approval by the institution's Institutional Review Board (IRB), a request was submitted to the university's IR office to obtain anonymous TCE raw data for all graduate courses in the College of Education across the two most recent fall semesters, as this is when the majority of research methods courses were taught at this institution.

2.3 Data Parsing

An exhaustive list of graduate course offerings and their course descriptions was generated. Courses that would be considered inappropriate for the present study, such as practicum courses, independent study courses, thesis/dissertation residency courses, and internship courses were removed from the sample frame. Remaining courses consisted of lecture, seminar, or laboratory sessions. Initially, the researchers wanted to collapse all courses into one of four categories: 1) quantitative; 2) qualitative; 3) other methods; 4) all other (non-methods) courses. Upon classifying data, only one qualitative course was identified for each semester. Due to concerns of anonymity, the researchers decided to include the qualitative course data into the "other methods" category. This resulted in a final classification schema consisting of three categories: 1) quantitative courses; 2) all other methods courses; and 3) all non-methods courses. This categorization proved ideal, as definitions for each category were straight forward (defined by curriculum), yielded appropriate sample sizes, and protected the anonymity of the data. In total, 15 distinct quantitative courses which included numerous lab sections (n=249), 7 other research methods (n=129) courses, and 146 non-methods courses (n=2,186) were sampled, resulting in a total sample size of 2,564 persons.

2.4 Analysis

Survey data were analyzed using a Rasch measurement model. In particular, the Rating Scale Model (Andrich, 1978) was selected because it is an appropriate model for polytomous data that share a common rating scale structure. According to the model the probability of a person *n* responding in category *x* to item *i*, is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_j)]} \qquad x = 0,1,...,m$$

where $\tau_o = 0$ so that $\exp \sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 1$, β_n is the person's position on the variable, δ_i is the scale value (difficulty to endorse) estimated for each item *i* and $\tau_1, \tau_2, \ldots, \tau_m$ are the *m* response thresholds estimated for the *m* + 1 rating categories.

Winsteps (Linacre, 2014) measurement software was used to perform the Rasch analysis. This included a calibration of person and item parameters, estimates of person and item reliability, indices of data to model fit, a principal components analysis of standardized residuals to test for multidimensionality, and other standard analyses.

Differential Item Functioning (DIF), an analysis technique based in item response theory, was used to compare calibrations across the three content categories. The goal of DIF is to ascertain if there are substantial differences amongst varying groups. If an item shows evidence of DIF it could be potentially biased (Holland & Thayer, 1986). Because a uniform instrument is administered to courses that vary significantly in terms of subject matter content, instructional methods, and other pedagogical issues, it is important to test the extent to which TCE items may potentially advantage or disadvantage some faculty teaching these courses. If it appears that numerous items are biased in a particular direction, one would have some empirical evidence to suggest that the instrument needs to be revised and improved and the inferences made about the findings reconsidered. For the present study, two iterations of DIF analyses were performed. The first analysis simply detected statistically significant differences among mean logit scores across categories at the .05 alpha level. Because compounding error (Type I) is a possibility when making multiple comparisons, a Bonferroni correction was applied to control for false positive detections in the second iteration of the DIF analysis.

3. Results

3.1 Validity and Reliability

Rasch models require data fit the model's expectations, thus we evaluated fit statistics for the overall model as well as for individual people and items. Infit and outfit statistics indicate where observed data in the response vector may not meet the model's expectations. Results indicate person mean square fit statistics (infit and outfit, respectively) were 1.01 and .98, and item mean square fit statistics were 1.00 and .99. These measures indicate very good fit, thus providing evidence for noise-free measures. Person and items fit statistics were evaluated and all items appeared sound. A couple hundred persons appeared to misfit the recommended range provided by Wright and Linacre (1994) of .6-1.4, however, because overall fit was so sound it was unnecessary to remove any data to improve overall fit.

Numerous quality control checks are necessary to ensure the data are sufficiently unidimensional for appropriate for a Rasch analysis. We investigated dimensionality by conducting a Principal Components Analysis of standardized residual correlations (PCA) that indicated 51.7% of the variance was explained. The eigenvalue for the first contrast was 2.6, representing 6.6% of the variance. This information suggests the data were sufficiently unidimensional for the purposes of measuring a single (and primary) latent trait. The strength of the secondary dimension is about 2 to 3 items. No other dimensions were detected.

Rasch models produce reliability estimates for both persons and items. Person reliability estimates ranged from .90 (real) to .91 (model), and item reliabilities were both .99 for real and model estimations. These estimates indicate the scores are highly reproducible.

3.2 DIF Results

DIF results were determined by comparing summative item level calibrations for each of the three groups. Results indicate a number of items were statistically significant at the .05 alpha level (indicated by "*"). A Bonferroni correction was applied to adjust for compounded error. This reduced alpha to .0026. Results that remained significant after the Bonferroni correction are denoted by "**". Table 2 illustrates results of the DIF analysis. Upon the flagging of statistically significant items we evaluated the directionality to determine which group may receive potentially biases reviews.

Table 2. Statistically Significant DIF Results

Items	Non-Methods v. Other Methods	Non-Methods v. Quantitative Methods	Other Methods v. Quantitative Methods
Q1			
Q2		*	*
Q3		*	
Q4		*	
Q5		**	*
Q6		**	*
Q7			
Q8			
Q9			
Q10			
Q11			
Q12			
Q13		*	
Q14		**	
Q15		*	
Q16		*	
Q17		*	
Q18			
Q19		*	

*indicates results were significant at .05 alpha.

**indicates results were significant after the Bonferroni corrected alpha of .0026.

When comparing non-methods courses to all other types of methods courses except those that were quantitative, no items were flagged to have evidence of DIF. When comparing non-methods to quantitative methods, 11 of 19 items were flagged for potential bias. Finally, comparing all other methods courses to quantitative courses, only 3 items were flagged.

4. Discussion and Conclusion

With regard to the directionality of these potential biases, let us first investigate non-methods compared to quantitative methods. Items 3, 4, 5, 6, 16 and 17 were more difficult for those in non-methods courses to endorse, and items 2, 13, 14, 15 and 19 were more difficult to endorse for those in quantitative courses. Upon investigating the content of these items, it appears students in non-methods courses were less likely to endorse items that were course-related in nature. Students in quantitative courses, however, were less likely to endorse items pertaining to the instructor. Two items were flagged for items pertaining to learning outcomes for each group. On the surface, it appears students perceived quantitative faculty as less likely to encourage class participation and less likely to stimulate students' interests in the subject matter. On the other hand, students were less likely to prefer the manner in which their non-methods courses were conducted (e.g., including helpful assignments, exams reflective of what was taught, fair and consistent grading, and distribution of assignments).

With regard to the three items that were significantly different between all other methods courses and quantitative courses, students in general methods courses were more likely to agree the textbook contributed to their understanding of the subject, but those in quantitative courses were more likely to agree that grading was fair and consistent and assignments were distributed fairly throughout the semester. These results reinforce the notion that students generally appreciated much of the manner in which quantitative courses were conducted. Given no significant differences were found between non-methods courses and all other research methods (non-quantitative) further illustrates that quantitative methods courses are particularly unique, and that the results drawn from TCEs pertaining to these courses may indeed be biased.

At many institutions, important decisions about promotion and tenure are linked to student ratings on TCEs. This study found that students' perceptions of faculty varied widely with regard to stimulating their interest in the subject and encouraging class participation. One may argue that these sentiments are of no fault to the instructor, as students simply may not have an interest in quantitative methods. This much is confirmed in impersonal conversations with students every day. However, items that speak to the quality of the instructor, such as his/her knowledge of the subject matter, ability to effectively present material, satisfactorily answering questions and being available outside of class revealed no meaningful differences. In many ways these types of items should be more telling of an instructor's competence than some of the other issues. Depending on how results are calculated and viewed (e.g., item level versus summative), there is reason to believe that students' internal biases may cause them to undermine effective teaching provided by instructors of quantitative courses. Decision-makers need to be keenly aware of this, and for the sake of fairness, investigate students' responses at the item level. This is critical because summative information obtained about an instructor may be misleading and inherently biased. When inferences from TCEs are used in a relatively high-stakes manner (for any instructor regardless of discipline), it is imperative that a fully-informed snapshot of results, ideally from multiples sources of evidence, is generated for decision-makers to draw appropriate inferences.

References

Achieve, Inc. (2010). On the Road to Implementation: Achieving the Promise of the Common Core State Standards. *Achieve, Inc*.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. http://dx.doi.org/10.1007/BF02293814

Banres, M.W., & Patterson, R. H. (1988) to accomplish the institutional mission. Paper presented at the *Annual Meeting of the Society for College and University Planning*, Toronto.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model. Fundamental measurement in the human sciences, 2nd edition*. Lawrence Erlbaum Associate.

Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability, 23*(4), 289-306. http://dx.doi.org/10.1007/s11092-011-9126-2

Cashin, W.E. 1990. Students do rate different academic fields differently. In Theall, M., and Franklin J., eds., *Student Ratings Of Instruction: Issues For Improving Practice*, Jossey-Bass, San Francisco. http://dx.doi.org/10.1002/tl.37219904310

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44,* 495-518. http://dx.doi.org/10.1023/A:1025492407752

Centra, J. A. (2009). Differences in responses to the student instructional report: Is it bias*? Listening, Learning, Leading,* 1-7.

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

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274. http://dx.doi.org/10.1037/0022-0663.97.2.268

Conaway, C. (2007). Supply and Demand of STEM Workers: STEM Jobs Are Growing, but Are Enough Massachusetts Students Qualified? Education Research Brief. Issue 2. *Massachusetts Department of Education.*

d' Apollina, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208. http://dx.doi.org/10.1037/0003-066X.52.11.1198

Dickman, A., Schwabe, A., Schmidt, J., Henken, R. (2009). Preparing the Future Workforce: Science, Technology, Engineering and Math (STEM) Policy in K-12 Education. *Public Policy Forum*.

Dunn, D. (2000). Letter Exchanges on Statistics and Research Methods: Writing, Responding, and Learning. *Teaching of Psychology, 27*(2), 128-130.

Earp, M. (2007). Development and Validation of the Statistics Anxiety Measure. (Doctorate of Philosophy Dissertation), University of Denver. (3279112)

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*(2), 199–242. http://dx.doi.org/10.1007/BF00976997

Francis, C. (2011). Student Course Evaluations: Association with Pre-Course Attitudes and Comparison of Business Courses in Social Science and Quantitative Topics. *North American Journal of Psychology, 13*(1), 141-154.

Franklin, J. & Theall, M. (1992). Disciplinary differences: Instructional goals and activities, measure of student performance, and student ratings of instruction. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Francisco, April.

Haladyna, T., & Hess, R.K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35*(6), 669-687. http://dx.doi.org/10.1007/BF02497081

Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences, 48*, 68-74. http://dx.doi.org/10.1016/j.paid.2008.02.021

Hendel, D. (1980). Experimental and affective correlates of math anxiety in adult women. *Psychology of Women Quarterly, 5*, 219-289. http://dx.doi.org/10.1111/j.1471-6402.1980.tb00958.x

Holland, P., & Thayer, D. (1986). Differential Item Performance and the Mantel-Haenszel Procedure. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00186.x

Jordan, J., & Haines, B. (2003). Fostering Quantitative Literacy. *Peer Review, 5*(4), 16-19. http://dx.doi.org/10.4135/9781483350578.n6

Kember, D., & Leung, D.Y.P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment and Evaluation in Higher Education, 33*(4), 341-353. http://dx.doi.org/10.1080/02602930701563070

Kember, D., & Leung, D.Y.P. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education, 52*(3), 278-299. http://dx.doi.org/10.1007/s11162-010-9194-z

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research* (Vol. 109, pp. 9-25): Jossey Bass, John Wiley & Sons, Inc. http://dx.doi.org/10.1002/ir.1

Linacre, J. M. (1987). An extension of the Rasch Model to multi-faceted situations. MESA University of Chicago.

Linacre, J. M. (2010). WINSTEPS® (Version 3.69.1). Computer Software. Beaverton, OR: Winsteps.com.

Macher, D., Paechter, M., Papousek, I., Ruggeri, K. (2012) Statistics Anxiety, Trait Anxiety, Learning Behavior, and Academic Performance. *European Journal of Psychological Education, 27,* 483-498. http://dx.doi.org/10.1007/s10212-011-0090-5

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–384). New York: Springer. http://dx.doi.org/10.1007/1-4020-5742-3_9

Marsh, H. W., & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issue of validity, bias, and utility. *American Psychologist, 52,* 1187-1197. http://dx.doi.org/10.1007/1-4020-5742-3_9

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754. http://dx.doi.org/10.1007/1-4020-5742-3_9

Mji, A., & Onwuegbuzie, A. (2004). Evidence of score reliability and validity of the Statistical Anxiety Rating Scale among technikon students in South Africa. *Measurement and Evaluation in Counseling and Development, 36*, 238-251.

Murtonen, M., Olkinuora, E., Tynjala, P., & Lehtinen, E. (2008). "Do I Need Research Skills in Working Life?": University Students' Motivation and Difficulties in Quantitative Methods Courses. *Higher Education: The International Journal of Higher Education and Educational Planning*, *56*(5), 599-612. http://dx.doi.org/10.1007/s10734-008-9113-9

Narayanan, W., Sawaya, W., Johnson, M. (2014). Analysis of Differences in Nonteaching Factors Influencing Student Evaluation of Teaching between Engineering and Business Classrooms. *Decision Sciences Journal of Innovative Education, 12*(3), 233-265. http://dx.doi.org/10.1111/dsji.12035

Neumann, L., & Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas. *Journal of Educational Psychology, 78*, 152-158.

Onwuegbuzie, A., & Wilson. (2003). Statistics anxiety: Nature, etiology, antecedents, effects and treatments. *Teaching in Higher Education, 8*, 195-209. http://dx.doi.org/10.1080/1356251032000052447

Papousek, I., Ruggeri, K., Macher, D., Paechter, M., Heene, M., Weiss, E.M. Schulter, G., Freudenthaler, H.H. (2012). Psychometric evaluation and experimental validation of the Statistics Anxiety Rating Scale. *Journal of Personality Assessment, 94*(1), 82-91. http://dx.doi.org/10.1080/00223891.2011.627959

Pritchard, R. E., & Potter, G. C. (2011). Adverse changes in faculty behavior resulting from use of student evaluations of teaching: A case study. *Journal of College Teaching and Learning, 8*(1), 1-7.

Ramsden, P. (1991) A performance indicator of teaching quality in higher education: The Course Experience Questionnaire, *Studies in Higher Education, 16,* 129–150. http://dx.doi.org/10.1080/03075079112331382944

Reason, R., Terenzini, P., & Domingo, R. (2006). First things first: Developing academic competence in the first year of college. *Research in Higher Education, 47*(2), 149-175. http://dx.doi.org/10.1007/s11162-005-8884-4

Rhodes, T. L. (2010). Quantitative Literacy VALUE Rubric Assessing Outcomes and Improving Achievement: Tips and tools for Using Rubrics: Association of American Colleges and Universities.

Richardson, F., & Woolfolk, R. (1980). *Mathematics Anxiety Test Anxiety: Theory, Research, and Applications*. Hillsdale, NJ.

Royal, K. D. (2010). Making meaningful measurement in survey research: A demonstration of the utility of the Rasch model. *IR Applications, 28*, 1-16.

Sawchuk, S. (2009). STEM Talent Increases, Jobs Decrease. *Education Week*, *29*(1), 1.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23,* 191–212. http://dx.doi.org/10.1080/0260293980230207

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.