Review of Educational Research
December 2013, Vol. 83, No. 4, pp. 598–642
DOI: 10.3102/0034654313496870

On the Validity of Student Evaluation of Teaching: The State of the Art

Pieter Spooren, Bert Brockx, and Dimitri Mortelmans
University of Antwerp

This article provides an extensive overview of the recent literature on student evaluation of teaching (SET) in higher education. The review is based on the SET meta-validation model, drawing upon research reports published in peer-reviewed journals since 2000. Through the lens of validity, we consider both the more traditional research themes in the field of SET (i.e., the dimensionality debate, the 'bias' question, and questionnaire design) and some recent trends in SET research, such as online SET and bias investigations into additional teacher personal characteristics. The review provides a clear idea of the state of the art with regard to research on SET, thus allowing researchers to formulate suggestions for future research. It is argued that SET remains a current yet delicate topic in higher education, as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical questions concerning the validity of SET.

Keywords: student evaluation of teaching, validity, higher education, educational policy

Student evaluation of teaching (SET) is used as a measure of teaching performance in almost every institution of higher education throughout the world (Zabaleta, 2007). Universities and university colleges have developed relatively complex procedures and instruments for collecting, analyzing, and interpreting these data as the dominant or, in some cases, the sole indicator of teaching quality. This widespread use is largely due to the apparent ease of collecting the data and presenting and interpreting the results (Penny, 2003). In addition, students are considered important stakeholders in the process of gathering insight into the quality of teaching in a course, as "the opinions of those who eat the dinner should be considered if we want to know how it tastes" (Seldin, 1993, p. 40). Although SET was originally intended primarily for formative purposes, such evaluations came into use for faculty personnel decisions in the 1970s (Galbraith, Merrill, & Kline, 2012). More recently, SET procedures have been included as a key mechanism in internal quality-assurance processes as a way of demonstrating an institution's performance in accounting and auditing practices (Johnson, 2000).

Purpose of SET

Student evaluation of teaching serves three purposes: (a) improving teaching quality, (b) providing input for appraisal exercises (e.g., tenure/promotion decisions), and (c) providing evidence for institutional accountability (e.g., demonstrating the presence of adequate procedures for ensuring teaching quality; Kember, Leung, & Kwan, 2002). In most institutions, SET is obviously used for formative purposes (e.g., as feedback for the improvement of teaching) as well as for summative purposes (e.g., mapping teaching competence for administrative decision-making and institutional audits; Arthur, 2009; Burden, 2008; Edström, 2008; Emery, Kramer, & Tian, 2003). These dual usages—and the unresolved tension between them (Penny, 2003)—makes the use of SET fragile. On the one hand, teachers are convinced of the value of SET as an instrument for feedback on their teaching (Balam & Shannon, 2010; Griffin, 2001; Kulik, 2001). Results obtained from SET help them to improve the quality of their teaching, as they provide instructors with insight into the strengths and weaknesses of their teaching practice, based on student opinions. For this reason, one can assume that many instructors welcome SET results in order to improve their subsequent teaching. On the other hand, it has been argued that the principal purpose of SET involves its use as a measure for quality monitoring, administrative policymaking (Penny & Coe, 2004), and for determining whether teachers have achieved a required standard in their teaching practice (Bolivar, 2000; Chen & Hoshower, 2003).

This justification for using SET in staff appraisals is related to an increasing focus on internal quality assurance and performance management in universities, which have become subject to the demands of consumer satisfaction (Blackmore, 2009; Olivares, 2003, Titus, 2008). Student satisfaction has come to play an important role in this *managerial approach* (Jauhiainen, Jauhiainen, & Laiho, 2009; Larsen, 2005; Valsan & Sproule, 2005), which is based on such key concepts as accountability, visibility, and transparency (Douglas & Douglas, 2006; Molesworth, Nixon, & Scullion, 2009). Teacher performance and the quality of teaching could thus be defined as the extent to which student expectations are met, thus equating student *opinions* with *knowledge*. For this reason, many faculty members have been questioning the validity and reliability of SET results for many years (Ory, 2001). Their concerns are comprehensible and appropriate as SET results can have serious effects on a teacher's professional career (Kogan, Schoenfeld-Tacher, & Helleyer, 2010).

Teachers' Concerns About the Validity and Reliability of SET

One of the major concerns involves the validity and the reliability of student opinions (i.e., the extent to which students are capable of providing appropriate teacher evaluations). Faculty concerns include the differences between the ways in which students and teachers perceive effective teaching, as well as the relationship of these perceptions to factors that are unrelated to good teaching. In some instances, SET surveys are even known as "happy forms" (Harvey, in Penny, 2003, p. 400) that are used for "personality contests" (Kulik, 2001, p. 10) or as a measure of "customer satisfaction" (Beecham, 2009, p. 135). Second, the sometimes poorly designed questionnaires suggest that the architects of the questionnaire also lack common understanding or consensus regarding what constitutes good or effective

teaching (Johnson, 2000; Knapper, 2001). In addition, many instruments are not tested with regard to their psychometric properties (Richardson, 2005).

Third, the common use of SET by means of administering standard questionnaires to be completed (in most cases, anonymously) by all students taking part in a course has been called into question. Administering SET in this way depersonalizes the individual relationship between teachers and their students. For example, Platt (1993) argued that "only the composite opinion of the majority of the students speaks" (p. 5) in a SET report, further warning each student that "you count only as you add to a sum into which you disappear without a trace" (p. 2). Most SET procedures allow little or no space for discussing, explaining, or negotiating the results with the students (Johnson, 2000). Fourth, the interpretation of SET results is more complicated than it looks, and it entails a risk of inappropriate use by both teachers and administrators for both formative and summative purposes (Franklin, 2001). Fifth, many faculty members are unaware of the sheer volume of research on SET (in which almost all of their concerns are addressed; Ory, 2001). It has been shown, however, that teachers who are familiar with the SET literature are more positive toward such evaluations (Franklin & Theall, in Paulsen, 2002). This lack of familiarity with the literature has generated a number of persistent myths or urban legends concerning SET, most of which have been invalidated in many research reports (Aleamoni, 1999).

Given these concerns, it is not that surprising that many teachers fear their next SET reports, even though they tend to see SET as useful for summative decision-making (Beran & Rokosh, 2009). In some cases, this leads to practices aimed at increasing SET scores rather than improving instruction (Simpson & Siguaw, 2000). The tyranny of the evaluation form may lead to grading leniency, which can result in grade inflation (Crumbley, Flinn, & Reichelt, 2010; Eiszler, 2002; Ellis, Burke, Lomire, & McCormack, 2003; Langbein, 2008; Oleinik, 2009; Redding, 1998). At the same time, many valuable thoughts and suggestions from students remain untouched, as faculty members who do not perceive SET instruments as valid measurements tend to ignore the results (Simpson & Siguaw, 2000).

Research on SET

As mentioned above, most stakeholders (e.g., teachers, students, administrators, and policymakers) are unaware of the number of research studies that have been conducted within the domain of SET. Several thousands of research studies have appeared since the publication of the first report on SET by Remmers and Brandenburg in 1927, addressing various elements of these evaluations. Nevertheless, the primary focus of these studies is on the validity of student opinions and their relationship to possible biasing factors (for overviews, see Aleamoni, 1999; Marsh, 1984, 1987, 2007b; Marsh & Roche, 1997; Wachtel, 1998). Although the majority of research shows that SET provides useful information to both teachers and administrators (Marsh, 1987; Ory, 2001; Penny, 2003), the validity of such evaluations continues to be called into question (Clayson, 2009).

Several authors (Olivares, 2003; Ory & Ryan, 2001; Onwuegbuzie, Daniel, & Collins, 2009) have developed conceptual validity frameworks for assessing the validity of SET (e.g., regarding the extent to which scores generated by SET instruments measure the variables they are intended to measure). These frameworks are based on Messick's (1989, 1995) unified conceptualization of validity. Onwuegbuzie

et al. (2009) developed a meta-validity model, which is subdivided to address construct, content, and criterion validity. Each of these types of validity is subdivided into areas of evidence. *Construct-related validity* (substantive validity, structural validity, convergent validity, discriminant validity, divergent validity, outcome validity, generalizability) addresses the extent to which an instrument can be seen as a meaningful measure of a given characteristic. *Content-related validity* (face validity, item validity, sampling validity) concerns the extent to which the items of an instrument are appropriate representations of the content being measured. *Criterion-related validity* (concurrent validity, predictive validity) is associated with the extent to which scores are related to another independent and external variable that can serve as a direct measure of the underlying characteristic.

The Current Study

The purpose of this article is to provide a systematic overview of the recent literature on SET (since 2000) using the meta-validity model for assessing the score validity of SET designed by Onwuegbuzie et al. (2009). Through this validity lens, we consider both the more traditional research themes in the field of SET (i.e., the dimensionality debate, the "bias" question, and questionnaire design) and some recent trends in SET research such as online SET and bias investigations into additional teacher personal characteristics. Our goal is to summarize the state of the art in SET research and provide a basis for developing ideas for future research.

Method

Literature Search

Given the inconsistent use of terminology concerning SET, the literature search for this study was based on a variety of terms that refer to the concept of SET (i.e., questionnaire-based student evaluations of an individual course). The following keywords were used (separately and in combination) when searching the electronic databases, Web of Science, EBSCO, and ERIC: *SET, student evaluation of teaching, student ratings, student ratings of instruction, teacher evaluation, teaching effectiveness, teaching performance, higher education*, and *student evaluations*. To ensure that the search would generate an overview of the state of the art in high-quality research concerning SET, the search was limited to articles published in international peer-reviewed journals since 2000. In the supporting texts, however, we will also discuss some classic studies published prior to 2000, which cannot be ignored.

We read the abstracts of 542 peer-reviewed journal articles. Each abstract was read by at least two authors to determine the article's relevance to the review (based on its relationship with validity issues regarding SET, methodology, and conclusions). The search was not limited to empirical studies but also included conceptual, theoretical, and review studies since such papers draw important conclusions for SET and SET research as well. The database search left us with 210 articles that were fully read by the first author. The snowball method was then used to identify additional works (including chapters in edited books) through the references listed in the selected articles.

For each article, specific information was noted, including (a) authors, (b) year of publication, (c) journal, (d) objectives of the study, (e) methodology, (f) important

findings and conclusions, and (g) relevance for this review (i.e., the validity of SET). In case of disagreements concerning important issues such as methodology, findings, and relevance to the review, an article was read by the other authors and discussed at a meeting. Based on the discussion, a decision was made to include or exclude that article. Although the literature search was limited to articles published in the English language, this review has an international character, as it includes 31 articles written by authors residing in 11 countries other than the United States.

The final database consisted of 160 pieces (158 journal articles and 2 book chapters), including empirical studies, theoretical pieces, and other types of articles. An initial reading of all articles suggested that each of the selected studies could be classified as addressing at least one of the aforementioned types of validity. The following sections provide a narrative review of the recent SET literature, organized according to the meta-validation model designed by Onwuegbuzie et al. (2009). In the reference list, all studies included in the review are indicated with an asterisk.

Results

Content-Related Validity

Sampling validity and item validity. Although SET has become common practice in many institutions, and although it has been the subject of thousands of research studies, there is a surprising amount of variation in the SET instruments used to collect feedback from students. The starting point seems simple: Institutions need instruments that will allow them to gather information (preferably comparable) for different types of courses as quickly as possible. Such surveys must also be highly economical (Braun & Leidner, 2009). Although Lattuca and Domagal-Goldman (2007) advocated the use of qualitative methods in SET, in practice, such evaluations usually consist of standardized questionnaires (including both rating scales and open-ended items) aimed at providing a descriptive summary of the responses for both the teacher and the teacher's department head, as well as the institution's educational board or personnel system (Richardson, 2005). Nevertheless, this dual objective has generated a panoply of SET instruments that vary greatly in both content and construction, due to the characteristics and desires of particular institutions. This variety has implications for the item validity (i.e., the extent to which SET items are decent representations of the content area) and the sampling validity (i.e., the extent to which the SET instrument as a whole represents the whole content area) of SET instruments.

Several well-designed and validated instruments are available, however, including the Instructional Development and Effectiveness Assessment (IDEA; Cashin & Perrin, 1978), the Students' Evaluation of Education Quality (SEEQ; Marsh, 1982; Marsh et al., 2009), the Course Experience Questionnaire (CEQ; Ramsden, 1991), the Student Instructional Report (SIR II; Centra, 1998), and the Student Perceptions of Teaching Effectiveness (SPTE; Burdsal & Bardo, 1986; Jackson et al., 1999), as well as the more recent Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS; Toland & De Ayala, 2005), the Student Course Experience Questionnaire (SCEQ; Ginns, Prosser, & Barrie, 2007), the Teaching Proficiency Item Pool (Barnes et al., 2008), the SET37 questionnaire for student evaluation of

teaching (SET 37, Mortelmans & Spooren, 2009), the Exemplary Teacher Course Questionnaire (ECTQ; Kember & Leung, 2008), and the Teaching Behavior Checklist (Keeley, Furr, & Buskist, 2010; Keeley, Smith, & Buskist, 2006). Validation procedures for other instruments have not been successful (Haladyna & Amrein-Beardsley, 2009).

Still, many instruments are developed without any clear theory of effective teaching (Ory & Ryan, 2001; Penny, 2003). They therefore lack any evidence of content validity and thus might fail to measure what they claim to measure (Onwuegbuzie et al., 2009). A clear understanding of effective teaching is a prerequisite for the construction of SET instruments. Although it is logical to assume that educational scientists have reached some level of consensus regarding the characteristics of *effective teachers* (e.g., subject knowledge, course organization, helpfulness, enthusiasm, feedback, interaction with students), existing SET instruments vary widely in the dimensions that they capture. In a theoretical article on the shortcomings of SET research, Penny (2003) argued in favor of establishing an interinstitutional task force to formulate a list of standards or characteristics within a common framework of effective teaching, which can be used as a basis for the development of SET instruments. We add two conditions: (a) institutions should be able to select the aspects that are most important, according to their educational vision and policy, thereby developing SET instruments that are consistent with their own preferences; and (b) all stakeholders (i.e., administrators, teachers, and students) should be involved in the definition of these characteristics.

Face validity. The latter condition is derived from the growing body of research showing that SET instruments, which are usually designed by administrators (based on some didactic model of teaching), do not always reflect the students' perspective concerning effective teaching. This disconnect affects the face validity of SET instruments (i.e., the extent to which the items of a SET instrument appear relevant to a respondent). For this reason, the results of such evaluations might be biased, as students tend to respond to items according to their own conceptions of good teaching (Kember, Jenkins, & Kwok, 2004). Kember and Wong (2000), for instance, concluded from interviews with 55 Hong Kong undergraduate students that students' perceptions of teaching quality should be seen as the result of an interplay between students' conceptions of learning (a continuum between active and passive learning) and students' beliefs about teaching of the lecturer (ranging between transmissive and nontraditional teaching). Besides, based on a sequential mixed-method analysis that led to a model that represented four meta-themes and nine themes that (according to 912 students) reflected students' conceptions of effective college teaching, Onwuegbuzie et al. (2007) concluded that three of these themes were not represented in the teaching-evaluation forms used at their university (student centered, expert, and enthusiast).

Bosshardt and Watts (2001) showed that, although the perceptions of students and teachers with regard to effective teaching are positively correlated, differences exist as well. For example, students care more about the teacher's preparation for class than instructors do. Pan et al. (2009) analyzed both quantitative (student ratings) and qualitative (students' comments in open-ended questions) student feedback data and found that, contrary to popular perception, students value the quality

of teaching (e.g., ability to explain, aiding understanding) more than they value particular instructor characteristics (e.g., humor, a charismatic personality, or storytelling skills). Barth (2008) concurred, having found that students' overall instructor ratings are driven primarily by the quality of instruction. Factor analysis and multiple regression analysis (167 classes, 30 instructors, +4,000 students) revealed that each of five factors (quality of instruction, course rigor, level of interest, grades, and instructor helpfulness) had a strong statistically significant relation with the overall instructor rating (with the five factors explaining 95% of the variance in the measure of overall instructor rating). Using multigroup SEM on a sample of 3,305 first-year and third-year undergraduate students in Hong Kong, Kember and Leung (2011) showed that students from four different disciplines (humanities, business, hard science, health sciences) shared the same ideas concerning the nature of an effective teaching and learning environment. There were nevertheless differences among disciplines concerning the extent to which some elements within this environment were brought into play. Pozo-Munoz, Rebolloso-Pacheco, and Fernandez-Ramirez (2000) used factor analysis based on data from a 39-item semantic differential scale to define the attributes of the *ideal teacher*, according to 2,221 students from a Spanish university. The most valued teacher characteristics were having knowledge, having adequate communication skills, and being competent in teaching.

Goldstein and Benassi (2006) noted that SET scores are higher when students and teachers agree on the characteristics of excellent lecturers. Based on a study that involved both students' and their teachers' conceptions of the *ideal teacher* and students' perceptions of teaching quality, they found that mean SET scores were higher (6.00 on a 7-point scale) in the no-discrepancy group (i.e., where students' and teachers' conceptions of the ideal teacher coincided) compared to the positive (when students rated the items on the ideal lecturer scale as more important than did their teacher) and negative (when teachers rated the items on the ideal lecturer scale as more important than did their students) discrepancy groups (mean SET scores were 5.52 and 5.68, respectively). ANOVA results showed a reliable quadratic effect (Cohen's $d = .26$) between the SET scores from these three groups. Kember and Leung (2008) derived nine principles of *good teaching* from interviews with award-winning teachers about their insights and practices. These principles form the basis for their Exemplary Teacher Course Questionnaire.

In summary, the research literature suggests that there is a risk that important SET stakeholders (i.e., teachers, students, and questionnaire architects) may differ in their conceptions with respect to effective teaching and, thus, should be involved in the process of defining good teaching, as well as in the design of SET instruments.

Construct-Related Validity

Structural validity and the dimensionality debate. Although it is widely accepted that SET should be considered multidimensional (given that teaching consists of many aspects) and that SET instruments should capture this multidimensionality, many authors and institutional boards argue in favor of single, global scores (Apodaca & Grad, 2005). Important questions thus arise with regard to the following: (a) the number and dimensions of effective teaching that can be distinguished and (b) the possibility of compiling an overall score based on these dimensions.

The SET literature reflects no consensus on the number and the nature of dimensions (Jackson et al., 1999). This lack of consensus is due to conceptual and methodological problems, given that (a) we lack a theoretical framework concerning effective teaching, (b) views on effective teaching differ both across and within institutions (Ghedin & Aquario, 2008), and (c) the measurement of dimensions continues to be relatively data-driven (with different post hoc analyzing techniques and different decision rules), with a few exceptions. The latter observation calls into question the structural validity of SET instruments (i.e., the extent to which the factors measured by a SET-instrument are consistent with the factor structure of the construct). Onwuegbuzie et al. (2009) argued that this method of assessing the dimensions of instruction does not guarantee that items included in SET forms represent effective teaching; instead, they should be seen as indicators of teaching performance (as perceived by the students).

Table 1 provides an overview of the dimensions captured in recently reported SET instruments, thereby demonstrating the wide variety that exists with regard to the aspects of teaching and course quality that are measured in SET. Feedback from students regarding particular aspects of courses is helpful as a guide for improving teaching. Teachers receive precise and detailed suggestions for refining their teaching in a particular course. Because SET is used for administrative decision-making as well, however, there is a need for a unidimensional and global SET score that provides a clear measure of overall teaching quality (McKone, 1999). In the 1990s, several leading SET authors entered into debate with regard to the dimensionality of SET. This debate also addressed the important question of whether SET scores on several dimensions could be captured by a single-order factor that represents a global construct (i.e., "general instructional skill") and whether such a factor could be used for summative purposes (see, e.g., Abrami & d'Apollonia, 1990, 1991; Marsh, 1991b; Marsh & Hovecar, 1991). The debate resulted in a compromise, which recommends the use of both specific dimensions and global measures for administrative decision-making, using the weighted averages of individual dimensions to generate an overall rating (Marsh, 1991a). Recent research provides further evidence on this matter. Many authors report evidence to support the multidimensionality of teaching, furnishing proof of higher order factors that reflect general teaching competency (Apodaca & Grad, 2005; Burdsal & Harrison, 2008; Cheung, 2000; Harrison, Douglas, & Burdsal, 2004; Mortelmans & Spooren, 2009).

Relationships between several dimensions of SET have been studied as well, using structural equation modeling. For example, Paswan and Young (2002) reported that the factors Course Organization and Student-Instructor Interaction have a positive effect on the factors Instructor Involvement (.66 and .78, respectively) and Student Interest (.60 and .65), on the 21-item Student Instructional Rating System (SIRS) instrument, whereas the factor Course Demands has a negative effect on these factors (–.38 and –.43). The authors argued that relationships between the factors in a SET instrument should be considered when interpreting the results. In a similar study, Marks (2000) reported that some constructs have large effects on others. For instance, students' ratings of teaching ability were affected by their expectations regarding the fairness of grading (.24). Marks concluded that SET may lack discriminant validity (see below) and advised caution when using global SET measures for summative decisions. Gursoy and Umbreit (2005) provided evidence for a model in which students' perceptions regarding the

TABLE 1
Summary of dimension numbers in SET instruments (ever since 2000)

Author	Instrument	N° of Dimensions	Dimensions
Barth (2008)	Institutional	5	Quality of instruction Course rigor Level of interest Grades Instructor helpfulness
Cohen (2005)	Institutional	2	Course Teacher
Ginns et al. (2007)	SCEQ	5	Good teaching Clear goals and standards Appropriate assessment Appropriate workload Generic skills
Gursoy & Umbreit (2005)	Institutional	4	Organization Workload Instruction Learning
Keeley et al. (2006) Keeley et al. (2010)	TBC	2	Caring and supportive Professional competency and Communicational skills
Kember & Leung (2008)	ETCQ	9	Understanding fundamental content Relevance Challenging beliefs Active learning Teacher–student relationships Motivation Organization Flexibility Assessment
Marks (2000)	Initial instrument	5	Organization Workload/difficulty Expected/fairness of grading Instructor liking/concern Perceived learning
Marsh et al. (2009) Marsh (1982) Coffey & Gibbs (2001)	SEEQ	9	Learning/value Instructor enthousiasm Organization/clarity Group interaction Individual rapport Breadth Exam/graded materials Readings/assignments Workload difficulty

(continued)

TABLE 1 (continued)

Author	Instrument	N° of Dimensions	Dimensions
Mortelmans & Spooren (2009)	SET37	12	Clarity of objectives
Spooren (2010)			Value of subject matter
			Build-up of subject matter
			Presentation skills
			Harmony organization course-learning
			Course materials
			Course difficulty
			Help of the teacher during the learning process
			Authenticity of the examination(s)
			Linking-up with foreknowledge
			Content validity of the examination(s)
			Formative evaluation(s)
Shevlin, Banyard, Davies, & Griffiths (2000)	Initial instrument	2	Lecturer ability
			Module attributes
Toland & De Ayala (2005)	SETERS	3	Instructor's Delivery of Course Information
			Teacher's Role in Facilitating Instructor/Student Interactions
			Instructor's Role in Regulating Students' Learning

Note. Keeley et al. (2006) found a good fit for one-factor model to the data as well. ETCQ = Exemplary Teacher Course Questionnaire; SCEQ = Student Course Experience Questionnaire; SEEQ = Students' Evaluation of Education Quality; SETERS = Students' Evaluation of Teaching Effectiveness Rating Scale; TBC = Teaching Behavior Checklist..

organization, course workload, and instructional abilities of their teachers have a positive impact on a fourth construct, their perception of learning (the estimated standardized path coefficients were .32, .04, and .60, respectively, $R^2 = .78$).

In summary, SET researchers agree that SET and SET instruments should capture multiple aspects (dimensions) of good teaching practice. Due to the absence of an agreement with respect to the number and the nature of these dimensions, which should be based on both the theory and empirical testing, SET instruments vary greatly in both the content and the number of dimensions. Additionally, recent research has revealed that many dimensions in SET instruments seem to be affected by a global (unidimensional) construct, which could be used for summative purposes. Thus, on the one hand, one could use the results on one or more particular dimensions when working on the improvement of (teaching) a course. On the other hand, an overall score derived from the (weighted) scores on dimensions of which it is known that they belong can be used to create a general factor representing

general teaching competency, which in turn can be used for the evaluation of teaching staff.

Convergent validity. The most common method for assessing the convergent validity of SET instruments is to examine the relationship of SET scores to student achievement (objective measure) or student perceptions of learning (subjective measure), which are considered proxies for the students' actual learning. Reviews and multisection studies suggest positive and moderate correlations between *student grades* and SET scores (Onwuegbuzie et al., 2009), varying between .10 and .47 (Cohen, 1981; Feldman, 1997). These studies also provide evidence regarding the criterion-related validity (concurrent validity) of SET.

Recent studies by Braun and Leidner (2009) and by Stapleton and Murkison (2001) indicate moderate to strong statistically significant associations between *students' self-reported acquisition of competence* and their satisfaction with teaching behavior. In these studies, correlation coefficients ranged between .28 and .75. Based on a meta-analysis of the literature (with a majority of the studies conducted in the 1970s), Clayson (2009) found a small average relationship (.13) between students' learning (i.e., *testing results*) and SET. Galbraith et al. (2012) suggested that the relationship between student achievement (as measured by a standardized learning-outcome test) and SET scores is nonlinear, with the most effective teachers falling within the middle percentiles of SET scores. Other researchers have found little or no support for the validity of SET as a predictor of student learning (e.g., Mohanty, Gretes, Flowers, Algozzine, & Spooner, 2005; Stark-Wroblewski, Ahlering, & Brill, 2007).

In this regard, however, it is appropriate to question the ways in which student achievement has been measured in previous studies. Student perceptions of learning might not always reflect actual learning (e.g., students could think that they had learned a lot during a course, even if they failed the examinations). And because student outcomes on objective tests are affected by other factors as well (e.g., prior knowledge, interest in the subject matter), they cannot be considered precise measures of actual student learning in a course. For this reason, a pretest is needed at the beginning of the course to estimate accurately how much learning individual students acquired at the end of the course. Students who are already familiar with the subject matter might receive good grades even though they do not learn very much, whereas slower students might fail the examinations even though they achieve considerable learning progress during the course. Future research using pretests and posttests of student achievement can provide useful insights into discussions of the relationship between student learning and SET.

Most authors agree that SET is correlated with teachers' self-evaluations, alumni ratings, and evaluations by trained observers (Marsh, 1987; Richardson, 2005; Roche & Marsh, 2000). This finding provides further evidence supporting the convergent validity of SET. Renaud and Murray (2005) reported a moderately strong correlation (.54) between SET and actual teaching behavior, as observed from videotapes. Given the relatively small correlations between SET and peer or administrator ratings, it is important to consider that SET is only one of many instruments available for mapping teaching effectiveness (Marsh & Roche, 1997). On many campuses, however, SET is used as an important (and, in some cases, the

sole) indicator of teaching quality in personnel decisions, implying that only one important stakeholder is involved in the evaluation process. Given the risk of differences among stakeholders regarding the concept of teaching effectiveness, and given that the persistent feelings of teachers that student evaluations may be biased by external characteristics, we argue that personnel files should include other measures of teaching quality (e.g., teachers' reflection on their SET scores, observation reports by peers or educational experts) as well.

Several authors (Burdsal & Harrison, 2008; Emery et al., 2003) also argue in favor of teaching portfolios, which contain various indicators of teaching performance, with student evaluations as one component. At the institutional level, SET can be included as one indicator (e.g., in addition to student progress and retention rates) when using DEA (Data Envelopment Analysis) to explore an institution's educational performance using the learning performance of its students (Montoneri, Lee, Lin, & Huang, 2011; Montoneri, Lin, Lee, & Huang, 2012).

In summary, the research literature revealed the existence of (small to strong) positive correlations between SET scores and student achievement, expert ratings of teaching behavior, self-ratings, and alumni ratings. These results provide evidence of the convergent validity of SET. However, due to the variety in stakeholders' views concerning good teaching and due to the variety in the measurement of student achievement, SET should not be the only indicator of teaching effectiveness in personnel files.

Discriminant validity and divergent validity. Many recent SET studies continue to address the question of bias, or the effect of factors that are not necessarily related to teaching quality on SET scores (Centra & Gaubatz, 2000). This issue involves the discriminant validity and divergent validity of SET, which has received considerable attention from researchers, administrators, and teachers. Although most leading SET researchers are convinced of the validity of SET, as research has found potentially biasing factors to be of little or no influence (Centra, 2003; Marsh & Roche, 2000), bias studies continue to play a central role in the recent literature.

Table 2 provides an overview of recent studies that address student-related, teacher-related, and course-related characteristics that might affect SET. Although it is not our intention to discuss each of these studies, it is clear that not all of the reported characteristics should be considered biasing factors. Some are meaningful indicators of student learning and are therefore logically related to effective teaching and SET. For example, student effort and class attendance indicate the interest and motivation of students in a particular course and are at least partly dependent upon the organization of and the teaching in that course. The experience, rank, and research productivity of the teacher are valuable indicators of a teacher's educational skills and knowledge of the subject matter.

On the other hand, although the course discipline and the sexual orientation of the teacher have nothing to do with effective teaching, they could be biasing factors for SET. The same applies to the teacher's gender or race. Further discussion concerns whether several other variables should be interpreted as biasing factors. For example, the relationship of SET to both course workload and student grade expectations continue to provoke discussions among SET researchers (for

TABLE 2
Relationships between student, teacher, and course characteristics and SET scores

Characteristic	Author(s)	Measure	Significant?	Interpretation
Student				
Student's cognitive background	Ting (2000)	Student's major and year of enrollment	Y	Mature students majoring in the same subject as the course, give higher SET
Class attendance	Beran & Violato (2005)	Frequency of attendance in the course	Y	Students who attend most classes (because of interest, motivation, being likely to learn, etc.) provide higher SET
	Davidovitch & Soen (2006a)		Y	
	Spooren (2010)		Y	
Students' effort	Heckert, Latier, Ringwald-Burton, & Drazen (2006)	Student effort (i.e., preparation for class, in-class behavior, etc.)	Y	Teachers who encourage students to make more effort, get higher SET
Expected grade	Beran & Violato (2005)	Student's expected grade	Y	The higher the expected grade, the higher SET
	Griffin (2004)		Y	
	Guinn & Vincent (2006)		Y	
	Langbein (2008)		Y	
	Maurer (2006)		Y	
	McPherson (2006)		Y	
	McPherson & Todd Jewell (2007)		Y	
	McPherson, Todd Jewell, & Kim (2009)		Y	
	Olivares (2001)		Y	
	Remedios & Lieberman (2008)		Y	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Stapleton & Murkison (2001)		Y	
	Marsh & Roche (2000)	Student's expected grade	Y	The higher the expected grade, the higher SET. But some SET factors are unrelated to expected grade, and relationship grade–SET is nonlinear (the highest grades are not correlated with SET)
	Isely & Singh (2005)	Expected grade at the class level	Y	SET are higher in classes in which students expect higher grades
	Centra (2003)	Student's expected grade	N	
	Stodnick & Rogers (2008)		N	
Final grades	Langbein (2008)	Student's final grade	Y	The higher the grade, the higher SET
	Spooren (2010)		Y	
Study success	Spooren (2010)	Passing the examinations in one or two times	Y	Students who had to retake the examinations for the course, give lower SET
Student's gender	Basow, Phelan, & Capotosto (2006)	Student's gender and teacher's gender	Y	There seem to be some gender preferences (i.e., female students give higher ratings to female teachers)
	Centra & Gaubatz (2000)			
	Kohn & Hartfield (2006)	Student's gender and teacher's gender	Y	Female students give higher SET than male students Female students give higher SET to male teachers than male students
	Santhanam & Hicks (2001)	Student's gender	Y	Female students give higher SET than male students
	Smith, Yoo, Farr, Salmon, & Miller (2007)		Y	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Spooren (2010)		N	
Student's goals	Remedios & Lieberman (2008)	Student's goal orientation (i.e., competitive, mastery, etc.)	Y	Students with a mastery goal are more likely to give positive SET
Student's age	Spooren (2010)	Students' age	Y	The greater the age, the higher SET
Grade discrepancy	Griffin (2004)	Difference between expected grade and believed deserved grade	Y	Students tend to punish teachers when expected grades are lower than they believed to deserve
Grading leniency	Griffin (2004)	Student's perception of instructor's grading	Y	The more lenient the grading, the higher SET
	Olivares (2001)		Y	
Pre-course interest	Olivares (2001)	Level of interest in the course	N	
Interest change during the course	Olivares (2001)	Interest change (increased, decreased, stable)	Y	Interest change during the course is positively associated with SET (increased interest leads to higher SET)
Precourse motivation	Griffin (2004)	Desire to take the course	Y	The stronger the desire to take the course, the higher SET
Teacher				
Instructor's gender	Basow & Montgomery (2005)	Teacher's gender	Y	Female teachers receive higher SET
	Smith et al. (2007)		Y	Female teachers receive higher SET
	McPherson et al. (2009)		Y	Male teachers receive higher SET
	McPherson & Todd Jewell (2007)		N	
Instructor's reputation	Griffin (2001)	Instructor reputation as perceived by the students	Y	Teachers with a positive reputation receive higher SET

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Research productivity	Stack (2003)	Citations and post-PHD year	Y	The better a teacher's quality of research, the higher SET
	Ting (2000)	Number of publications	N	
Instructor's teaching experience	McPherson et al. (2009)	Total semesters of teaching experience	Y	More experienced teachers receive higher SET
	McPherson & Todd Jewell (2007)		Y	
	McPherson (2006)	Teaching experience (<5, 5–10, 11+ semesters)	Y	
Instructor's age	McPherson et al. (2009)	Teacher's age	Y	Younger teachers receive higher SET
	Spooren (2010)		N	
Instructor's language background	Ogier (2005)	English as a second language (ELS) vs. native speakers	Y	ELS speakers receive lower SET than native speakers (especially in the science faculties)
Instructor's race	McPherson et al. (2009)	Teacher's race	Y	White teachers receive higher SET in upper-level courses
	McPherson & Todd Jewell (2007)		Y	
Instructor's tenure	McPherson & Todd Jewell (2007)	Tenured vs. nontenured faculty	Y	Nontenured faculty receive lower SET
Instructor's rank	McPherson et al. (2009)	Adjunct instructors vs. tenure-track faculty	Y	Adjunct instructors receive higher SET than tenure-track faculty
	Spooren (2010)		Y	
	Ting (2000)	Full professors vs. professors, associate professors, lecturers, and junior lecturers	N	(Full) professors receive higher SET than associate professors and professors
		Senior lecturers vs. all other ranks		

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Instructor's sexual orientation	Ewing, Stukas, & Sheehan (2003)	Sexual orientation (gay/lesbian vs. unspecified)	Y	After *strong* lectures, known gay/male teachers receive lower SET, but after *weak* lectures they receive higher SET
Instructor's personal traits	Shevlin et al. (2000)	Teacher charisma	Y	A modeled "charisma" factor explains 69% and 37% of the variation in the "lecturer ability" and "module attributes" factors, respectively
	Clayson & Sheffet (2006)	Teacher personality (Big Five)	Y	Students' evaluations of their instructor's personality (Big Five) show significant correlations with SET
	Patrick (2011)		Y	
	Campbell, Gerdes, & Steiner (2005)	Physical attractiveness	N	
	Feeley (2002)		Y	Measures of instructor physical attractiveness have significant relationships with measures of effective teaching
	Gurung & Vespia (2007)		Y	Likable, good-looking, well-dressed, and approachable teachers receive higher SET
	Hamermesch & Parker (2005)		Y	Good-looking teachers receive higher SET (besides, the impact is larger for male than for female instructors)
	Riniolo, Johnson, Sherman, & Misso (2006)		Y	Professors perceived as attractive received student evaluations about 0.8 of a point higher on a 5-point scale
	Wendorf & Alexander (2005)	Instructor fairness	Y	SET is significantly related to perceptions of the fairness of grading procedures, the fairness of instructor–student interactions, and the fairness of the expected grades

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Kim, Damewood, & Hodge (2000)	Professor attitude	Y	Instructors who are perceived as approachable, respectful, pleasant … receive higher SET
	Dunegan & Hrivnak (2003)	Image compatibility	Y	SET scores are significantly related to image compatibility (i.e., the comparison between an image of an "ideal" instructor with an image of the instructor in this course)
	Delucchi (2000)	Instructor likability	Y	Instructors who are rated high in likability receive higher SET
	Tom, Tong, & Hesse (2010)	Initial impressions of a teacher	Y	SET based upon 30-s video clips of instructors in the classroom correlate strongly with end of the term SET
Course				
Class size	Bedard & Kuhn (2008)	Class size	Y	Nonlinear, negative relationship between class size and SET (relationship becomes stronger for higher class sizes)
	McPherson (2006)		Y	
	McPherson et al. (2009)		N	
	Ting (2000)		N	Negative relationship between class size and SET
Class attendance rate	Ting (2000)	Class attendance rate (ratio of students present in evaluation exercise and the class size)	Y	The higher the class attendance rate, the higher SET
Class heterogeneity	Ting (2000)	Index of diversity (based on students' years of enrolment in the same class)	N	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Course difficulty	Remedios & Lieberman (2008)	Student's perceived course difficulty	Y	The more difficult the course, the lower SET
	Ting (2000)	Identified by institution	N	
Course discipline	Basow & Montgomery (2005)	Course discipline	Y	Natural science courses receive lower SET
	Beran & Violato (2005)		Y	Natural science courses receive lower SET
Course workload	Centra (2003)	Student's perception of course workload	Y	SET are lower for both difficult and too elementary courses; "just right" courses receive the highest SET
	Marsh & Roche (2000)		Y	
	Marsh (2001)		Y	
	Dee (2007)		N	A positive relationship between course workload and SET A positive, nonlinear relationship between good (useful) workload and SET (relationship becomes smaller for higher workloads)
Course level	Santhanam & Hicks (2001)	Course's year level	Y	SET in higher year level are more positive
Course type	Beran & Violato (2005)	Lab-type vs. lectures/tutorials	Y	Lab-type courses receive higher SET
Elective vs. required courses	Ting (2000)	Required vs. elective courses	Y	Elective courses receive higher SET (lecturing performance)
General education vs. specific education	Ting (2000)	General vs. specific course contents	Y	Courses with specific content matters receive higher SET
Syllabus tone	Harnish & Bridges (2011)	Friendly vs. unfriendly syllabus tone	Y	Teachers with a friendly written syllabus tone receive higher SET

overviews, see Brockx, Spooren, & Mortelmans, 2011; Griffin, 2004; Gump, 2007; Marsh, 2001, 2007b). Many SET studies provide evidence to support the validity hypothesis with regard to interpreting the relationship between expected grades and SET, thus suggesting that the positive relationship between expected grades and SET has to do with the fact that students who have learned a great deal—and who thus expect good grades—assign higher SET scores for their teachers. Such studies have also rejected the hypothesis concerning the existence of a negative (and thus biasing) relationship between course workload and SET (Marsh, 2001; Marsh & Roche, 2000). Nevertheless, other authors continue to advocate the grading-leniency hypothesis (i.e., teachers can *buy* good evaluations by giving high grades; see, e.g., Isely & Singh, 2005; Langbein, 2008; McPherson, 2006; McPherson & Todd Jewell, 2007), drawing upon attribution theories and measures of the instructor's grading leniency (as perceived by students) to support their argument (Griffin, 2004; Olivares, 2001).

In addition to research on the impact of the classic and potentially biasing factors, a considerable amount of research focuses on the impact of psychological dynamics on SET. First, some authors argue for the possibility of halo effects in SET. A halo effect can be understood as "a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behaviour" (Feeley, 2002, p. 226). The contention is that students base their evaluations of a given teacher or course on a single characteristic of that teacher or course, subsequently generalizing their feelings about this characteristic to most or all other unrelated characteristics of the teacher or course. Shevlin et al. (2000) defined a charisma factor that explains a large portion of the variance in several factors (69% and .39% in the factors Lecturer Ability and Module Attributes, respectively) included in their SET instrument. Significant correlations (ranging between .28 and .72) have been observed among all measures in the SET instrument developed by Feeley (2002), which also includes irrelevant measures (e.g., physical attractiveness).

Ever since Ambady and Rosenthal (1993) found that students' opinions about teachers are formed within seconds of being exposed to the nonverbal behavior and physical attractiveness of these teachers, bias studies have also focused on other personal traits that are considered strongly related to SET. Examples include teacher personality as measured by the Big Five personality traits (Clayson & Sheffet, 2006; Patrick, 2011), physical attractiveness (Campbell et al., 2005; Gurung & Vespia, 2007; Hamermesch & Parker, 2005; Riniolo et al., 2006), instructor fairness (Wendorf & Alexander, 2005), professor attitude (Kim et al., 2000), image compatibility (Dunegan & Hrivnak, 2003), instructor likability (Delucchi, 2000), and initial impressions of a teacher (Tom et al., 2010).

Generalizability. Most of the contradictory research results on SET are due to the great variety of methods, measures, controlling variables, SET instruments, and populations used in these studies. This high degree of variation calls the generalizability of these results into question and makes it almost impossible to make statements concerning, for example, the global effect size of the concurrent validity coefficients with student achievement, or the strength of the relationship of various possibly biasing effects on SET scores. However, several researchers have found

that the effect of the possibly biasing factors on SET is relatively small. For instance, Beran and Violato (2005) found that various students and characteristics explained only 7% of the total variance in SET scores. Spooren (2010) reported small local effect sizes of 6.3% for students' grades and of 1.6% for the examination wherein the course grade was given (students that had to retake examinations give lower SET) on SET. The PRV (proportional reduction in variance statistic) for other student, course, and teacher characteristics was estimated close to 0. Smith et al. (2007) noted statistically significant effects of sex of students and sex of instructors on SET scores, but these predictors did not account for more than 1% of the explained variance in SET. These findings suggest that SET outcomes depend primarily upon teaching behavior (Barth, 2008; Greimel-Fuhrmann & Geyer, 2003).

Nevertheless, some authors recommend adjusting raw SET scores in order to purge them of any known biasing effect, especially when these results are used for ranking (McPherson, 2006; McPherson et al., 2009; Santhanam & Hicks, 2001). In this regard, future SET research could also explore the simultaneous administration of SET and such measures as the Marlowe-Crowne Social Desirability Bias Index (Crowne & Marlowe, 1960). This strategy might improve the adequacy of SET for making evaluative decisions, as it would allow the elimination of one type of bias from analyses.

Substantive validity. One crucial topic in the debate on the construct-related validity of SET concerns student behavior when completing SET questionnaires. This issue affects the substantive validity of SET instruments (i.e., the extent to which an instrument is consistent with the knowledge, skills, and processes that underlie a respondent's scores). Understanding how students react to certain questions (or types of questions) and being aware of response patterns provide information that could be useful in the construction of SET items and could increase the substantive validity of SET scores. Recent research has paid considerable attention to what should and should not be done when developing SET questionnaires that take into account the knowledge and skills supposed to underlie students' SET scores.

Instruments used in SET measure students' attitudes toward effective teaching, which should be seen as a latent construct. Such a construct is not immediately observable using a single-item approach that, although sometimes resulting in highly stable estimates (Ginns & Barrie, 2004), assumes that all aspects or dimensions of teaching quality can be observed unequivocally. Spooren, Mortelmans, and Denekens (2007) argued in favor of using Likert-type scales in which sets of items measure several dimensions of teaching quality. These scales allow a straightforward quality check (e.g., by calculating alpha statistics) for each dimension contained in a SET report. Multiple-item scales also provide both the administrator and the teacher with information on score reliability for each particular course evaluation.

Most SET instruments use Likert-type scales to gather information on the quality of teaching in particular courses. This choice is related to ease of use (for both administrators and teachers), given that scales grant a quick and clear view of student opinions regarding the teaching in a particular course. As many authors have observed, however, SET results are subject to bias due to both the content and

the structure of these scales. For example, Onwuegbuzie et al. (2009) cautioned questionnaire designers about using midpoint or neutral categories in SET scales. Based on several studies, Onwuegbuzie and Weems (2004; Weems & Onwuegbuzie, 2001) argued that the inclusion of a midpoint option attenuates the internal consistency of SET scores. Sedlmeier (2006) suggested that the way in which rating scales are constructed may also have an impact on SET scores. Sedlmeier's study addresses the effects of three types of scales: (a) endpoint numbering (uni-polar vs. bipolar scales), (b) different ranges in scales, and (c) the ordering of choices. Two of these effects (endpoint numbering and different ranges in scales) are quite substantial and should therefore be considered when constructing SET questionnaires.

Robertson (2004) concluded that SET scores can be affected by item saliency and the position of questions in the questionnaire. Moreover, the SET scores observed in that study improved when students were asked to provide explanations for their answers. With regard to the number of response options, Landrum and Braitman (2008) reported that students use a greater range of points on a 5-point scale than they do on a 10-point scale. Students are more accurate using a 5-point scale, as it is easier to differentiate between five options than it is to distinguish between 10. In a study of response patterns in SET forms, Darby (2008) found that students tend to respond at the favorable end of evaluation scales, which does not mean that all courses were—in actual fact—good. For this reason, Darby argued that SET reports should include a means of comparison by, for instance, asking students to rank a course in comparison to other courses.

Recent SET studies have also focused on acquiescence (yea saying) as a response style, although the results have been mixed. Although a recent study (Spooren, Mortelmans, & Thijssen, 2012) yielded no evidence of acquiescence in SET scores, Richardson (2012) identified both acquiescence and extreme responding as consistent traits in SET. The precise impact of these traits remains unclear, but caution is advised with regard to possible bias due to acquiescence and extreme responding in SET results. To this end, studies by Dolnicar and Grun (2009) and Spooren et al. (2012) provide lists of recommendations for avoiding, controlling, and correcting for acquiescence and extreme responding in SET. These lists include using semibalanced scales, calculating reliability estimates, counting frequencies, comparing groups of students, and employing such correction methods as the subtraction of individual means and division by the individual standard deviations.

Acquiescence sometimes results from excessively demanding SET practices in many institutions, which overburden students with evaluations. Sampling therefore appears to be an efficient strategy that does not decrease the validity and reliability of the results (Kreiter & Laksham, 2005). Roszkowski and Soven (2010) argued against the use of balanced scales and advocated using only positively worded items in SET questionnaires. In their opinion, the use of bi-directional (i.e., positive and negative) item wording produces ambiguous results, due to carelessness on the part of students. Given that response patterns might also emerge from poor item wording (e.g., vague, unclear, too difficult, irrelevant), attention should be paid to the formulation of items. Based on think-aloud interviews with students, Billings-Gagliardi, Barrett, and Mazor (2004) observed that students understand educational terms in different ways and therefore make different judgments.

Validity of Student Evaluation

Although many studies have been conducted on the reliability, validity, and utility of scales, most SET forms also include open-ended questions. Students are invited to share more specific opinions and suggestions concerning both the course and the teacher. In an analysis of written comments from students, Nasser and Fresko (2009) observed that such comments are more often positive (59% positive units and 41% negative units) and general (rather than specific), and that they correlate with answers to the closed-ended questions in the questionnaire, as well as to specific characteristics of the course (correlations ranged between .23 and .57). The latter suggests that both closed-ended and open-ended questions should be included in SET forms, as written comments allow students to explain the scores that they assign for closed-ended items and to draw attention to topics that were not addressed in the closed-ended part of the form.

All of these findings suggest that questionnaire designers should be aware of the consequences of their choices (single item vs. Likert-type scale approach, the number of options, midpoint options) when constructing SET items, since their substantive validity is at risk. Response patterns, neutral responses, favorable answers, and different conceptions concerning educational terms might greatly influence SET scores.

Outcome validity. The previous sections demonstrate that SET scores may be affected by the instruments used, as well as by the opinions of student, perhaps even to the point of challenging their validity. Furthermore, much of the existing SET literature focuses on these topics. Even if all of these biasing challenges are under control, however, and even if SET provides valid information concerning the quality of teaching, it is still possible for such evaluations to be administered and used in inappropriate ways. Use affects the outcome validity of SET. Onwuegbuzie et al. (2009) argued that evidence concerning the outcome validity of SET may be the weakest of all evidence regarding validity issues.

Penny (2003) stated that the ways in which administrators engage with SET constitute one of the greatest threats to the validity of SET. Although guidelines for the collection and interpretation of SET data are available, many SET users are not sufficiently trained to handle these data, and they may even be unaware of their own ignorance. Moreover, they lack knowledge about the existing research literature on SET. Although the misuse and mis-collection of data might have consequences for both the improvement of teaching and the careers of the teachers involved, little research is available concerning this topic. In this section, we provide an overview of recent SET research concerning the collection and interpretation of SET data, which focuses primarily on attitudes toward SET and the relationship between SET and the improvement of teaching.

Students' attitudes toward SET. Students' attitudes toward the goals of SET are apparently important when collecting SET. If students see no connection between their efforts in completing SET questionnaires and the outcomes of these evaluations (e.g., teacher awards or improvements in teaching or course organization), such evaluations may become yet another routine task, thus leading to mindless evaluation behavior (Dunegan & Hrivnak, 2003). Spencer and Schmelkin (2002) reported from a mail survey to a random sample of students that students are

generally willing to participate in SET procedures, and that they do not fear possible repercussions for giving negative evaluations (the mean scores on the factors Reluctance to Do Evaluations and Potential Repercussion Against Students, as measured by means of a 7-point scale with 1 as *disagree very strongly*, were 2.94 and 2.24, respectively). Nevertheless, they have little confidence that their evaluations are actually taken into account by either administrators or teachers (the mean scores on the factors Impact of Teaching on Students and Student Opinion Taken Seriously, as measured by means of a 7-point scale with 1 as *disagree very strongly*, were 4.55 and 4.28, respectively).

Students are also ambivalent about the relative utility of the SET process. Chen and Hoshower (2003) observed that, according to the students, providing feedback for the improvement of teaching is the most attractive outcome of a teaching-evaluation system. The expectations that students have concerning this outcome have a significant impact on their motivation to participate in evaluations. This is an important finding, as response rates in SET are generally low (fluctuating between 30% and 50%), especially in the case of online course evaluations (Arnold, 2009; Dommeyer, Baum, Hanna, & Chapman, 2004; Layne, Decristoforo, & McGinty, 1999), and might affect SET scores (McPherson, 2006).

With regard to their use of SET, students reported that they find SET somewhat useful (e.g., for course selection), although there is variation according to frequency of use, as well as according to student and program characteristics (Beran, Violato, Kline, & Frideres, 2009). Although students are more likely to choose courses that have good SET results (if they are available), the possibility of acquiring useful knowledge remains the most important selection criterion (Howell & Symbaluk, 2001; Wilhelm, 2004). Using SET for administrative decision-making was not found to be an important motivator for student participation in SET (Chen & Hoshower, 2003).

Teachers' attitudes toward SET. Teachers' attitudes toward SET are important for both the collection and the use of SET, given that the usefulness of these evaluations for the improvement of teaching depends upon the extent to which teachers respond to and use them (Ballantyne, Borthwick, & Packer, 2000). Nevertheless, Moore and Kuol (2005a) argued that surprisingly few studies examine faculty perceptions and the nature of teacher reaction to student feedback. Moore and Kuol (2005b) developed a tentative quadrant for understanding teacher reactions to SET (i.e., endorsement, ego protection, problem solving, and repair), based on a comparison of positive/negative self-evaluations with positive/negative SET. The authors observed two risks related to these reactions: fixation on minor issues (e.g., making changes to the layout of a PowerPoint presentation) and de-motivation, dejection, and withdrawal from the commitment to teaching effectiveness. Yao and Grady (2005) found from interviews with 10 faculty members that teachers care about feedback from students, although they experience anxiety and tension concerning the summative purposes of SET.

The ways in which teachers use SET varies according to background and experience. Arguing that responding to feedback is indeed a complex process, Arthur (2009) developed a typology of factors (e.g., personality, student characteristics, teaching and learning strategies) that affect teachers' individual responses to

negative feedback (i.e., tame, blame, reframe, shame). Understanding the ways in which instructors respond to SET could help to overcome the doubts that teachers have regarding the validity of SET as an indicator of teaching quality, as well as their differing perceptions regarding the accuracy of SET (Simpson & Siguaw, 2000).

In general, teachers tend to agree that SET is an acceptable means of assessing institutional integrity and that it may be useful for administrative decision-making. Beran and Rokosh (2009) reported from a survey to 262 university teachers that 84% of the respondents support the use of SET in general, and that 62% of the respondents feel that department heads and deans make proper use of SET reports. Gender differences can be observed in perceptions of SET, however, with SET apparently having a greater negative impact on female teachers as they report a strong or moderate impact more often than male teachers when asked, "How much impact do you think your gender has on their evaluation of you?" (Kogan, Schoenfeld-Tacher, & Helleyer, 2010).

Based on interviews with 22 teachers, Burden (2008, 2010) observed a common recognition of the importance of SET. Nevertheless, only four of the teachers interviewed reported seeing the teacher feedback provided by SET as amounting to little more than hints and tips, as the evaluations did not reflect their perceptions of good teaching. The results of this study are supported by quantitative research results. Nasser and Fresko (2002) found from a survey with 101 instructors at a teacher's college that instructors consider SET of little value for the improvement of their teaching, and that teachers make little or no use of student feedback. In the above-mentioned study, Beran and Rokosh (2009) found that SET results are used for improving general teaching quality (57%), for refining overall instruction (58%), and for improving lectures (54%). SET results are least often used for specific changes in particular courses, such as textbooks (23%), examinations (24%), student assignments (28%), support materials (34%), or for refining instructional objectives (40%).

Adminstrators' attitudes toward SET. Although we are not aware of any recent study that include administrators' attitudes toward SET, it is reasonable to expect that they would be more positive with regard to the use and validity of such evaluations, as they provide a quick and easy indicator of teaching performance (Sproule, 2000). Nevertheless, administrators have challenged the validity of SET based on limited psychometric knowledge (Franklin, 2001; Sproule, 2000; Wolfer & Johnson, 2003). Administrators prefer aggregated and overall measures of student satisfaction, often failing to consider both basic statistical and methodological matters (e.g., response rate, score distribution, sample size) when interpreting SET (Gray & Bergmann, 2003; Menges, 2000) and making spurious inferences based on these data. For example, Franklin (2001) reported that about half of the SET administrators involved in the study were unable to provide sound answers to several basic statistical questions. The proper collection and interpretation of SET data depend upon administrators having sound methodological training and regular briefing on the major findings and trends in the research field.

SET and the improvement of teaching. An important outcome of SET would be, as mentioned above, to provide student feedback for the improvement of teaching in

particular courses. In the previous paragraph, we argued that many teachers do not find SET very helpful for such formative purposes and that they tend to ignore the comments and suggestions that students provide. These findings suggest that SET ultimately does not achieve the goal of providing useful information to an important stakeholder, with the ultimate goal of improvement. One important question addressed in the recent SET literature, therefore, involves the relationship between SET and the improvement of teaching. Davidovitch and Soen (2006b) showed that SET improves over time (with the age and seniority of teachers as particularly important predictors). Contrary to these results, however, a study by Kember et al. (2002) based on multiyear SET data from one university revealed no evidence that such evaluations contribute to the improvement of teaching, as SET scores did not increase over the years. These findings could be explained by several factors, including the organization and goals of SET in particular institutions, as well as the quality of the instruments and procedures that are used.

Consultative feedback on SET. Another possible explanation is that the student feedback obtained from the questionnaire is not used effectively. Marsh (2007a) concurred, saying that student feedback alone is not sufficient to achieve improvement in teaching. Using a multilevel growth-modeling approach, Marsh (2007a) demonstrated that SET reports are highly stable over time, including with regard to the individual differences between teachers. It is therefore important for teachers to have the opportunity to consult with colleagues or educational experts about their SET reports. In a longitudinal study, Dresel and Rindermann (2011) observed that consulting with faculty about their SET has a moderate to large positive effect (.68) on teaching quality, even when controlling for variables reflecting bias and unfairness. Lang and Kersting (2007) found that providing feedback by SET reports alone (without consultation) is far less effective than many assume in the long run. They noted a strong increase in SET results the next semester, which was followed by declines over the next three semesters.

Nasser and Fresko (2001) provided a typology of teachers who seek voluntary peer consultation regarding their SET reports. Three attributes were associated with this form of help seeking: lack of prior teacher training, teaching lecture courses, and being female. In addition, instructors were satisfied with their consultations, although they subsequently made few changes in their teaching. Relatedly, a meta-analysis by Penny and Coe (2004) on the effectiveness of consultation on student feedback showed that not all consultation practices are effective in improving teaching effectiveness. Consultative feedback should consist of more than simply interpreting the results and providing advice for teaching improvement. These authors listed eight strategies that are important when providing consultative feedback: (a) active involvement of teachers in the learning process, (b) use of multiple sources of information, (c) interaction with peers, (d) sufficient time for dialogue and interaction, (e) use of teacher self-ratings, (f) use of high-quality feedback information, (g) examination of conceptions of teaching, and (h) setting of improvement goals.

Predicting SET. Another strategy involves highlighting the discrepancy between predicted and actual ratings, which, according to Nasser and Fresko (2006), can

serve as an impetus for teaching improvement. According to these authors, teachers are generally quite good at predicting their SET scores. Nevertheless, the results revealed a trend in which teachers with lower ratings tend to overestimate their SET, and those with higher ratings tend to underestimate their SET (effect sizes of significant differences based on *t* tests between teachers' predictions and SET results ranged between .61 and 1.30). It is clear that all of these strategies lead to the inclusion of SET in a more holistic approach that stimulates teachers to be and remain to be reflective practitioners concerning their teaching, instead of merely taking note of the next SET report.

Criterion-Related Validity

As mentioned above, SET research reveals moderate to large positive correlations between SET scores and other indicators of teaching quality (e.g., student achievement, alumni ratings, self-ratings). These coefficients provide strong evidence for the concurrent and predictive validity of SET instruments' scores. In recent years, however, electronic evaluation appears to have replaced the classic paper-and-pencil questionnaire as the most common means of gathering SET in institutions throughout the world (Arnold, 2009; Nulty, 2008). Recent research has examined the validity of SET results that are obtained from such electronic procedures to ascertain if these procedures provide SET scores that are comparable to those obtained from the more classic paper-and-pencil procedures. In this section, we discuss research results that focus on the relationship between paper-and-pencil SET procedures and electronic SET procedures. Second, we consider the rise of online SET platforms (such as RateMyProfessors) and their relationship with SET scores obtained from institutional procedures.

Concurrent validity of electronic versus paper-and-pencil SET procedures. The primary reasons given for shifting to electronic SET include the following: (a) greater accessibility to students, (b) quick and accurate feedback, (c) no disruption of class time, (d) more accurate analysis of the data, (e) better written comments, (f) guaranteed student anonymity (e.g., decreased risk of recognition due to handwriting), (g) decreased vulnerability to faculty influence, (h) lower costs, and (i) reduced time demands for administrators (Anderson, Cain, & Bird, 2005; Ballantyne, 2003; Bothell & Henderson, 2003; Bullock, 2003; Tucker, Jones, Straker, & Cole, 2003). Some parties nevertheless fear that SET results obtained in this way are easier to trace and can be consulted by almost everyone (Gamliel & Davidovitz, 2005).

Moreover, response rates in such evaluation procedures are lower than is the case with paper-and-pencil questionnaires (Gamliel & Davidovitz, 2005). Dommeyer et al. (2004) reported average response rates of 70% for in-class surveys and 29% for online surveys. Johnson (2003) suggested several strategies for increasing electronic SET response rates, including encouragement by the faculty (i.e., if faculty members show genuine interest in SET, students will be more motivated to participate) and increasing the intrinsic motivation of students to participate (e.g., by highlighting their important role as raters), providing access to the electronic evaluation system, and clear instructions concerning participation in the SET process.

Several studies have investigated whether the shift toward electronic evaluations has affected SET scores. Studies by Leung and Kember (2005) and by Liu (2006) revealed no significant differences between SET scores obtained from paper-and-pencil evaluations and those obtained through electronic evaluations. These results support the concurrent validity of both types of instruments, although Venette, Sellnow, and McIntyre (2010) reported that student comments in electronic evaluations are more detailed than are those in paper-and-pencil questionnaires. At the aggregate level, Barkhi and Williams (2010) noted that electronic SET scores are lower than are those obtained with paper-and-pencil surveys. These differences disappear, however, when controlling for course and instructor. Moreover, electronic SET instruments generate more extreme negative responses to Likert-type items than do paper-based surveys. Paper-and-pencil questionnaires have traditionally been administered during the last class of a particular course, thus making them subject to little or no influence from the examination for that course. In contrast, Arnold (2009) identified differences in SET scores obtained in electronic surveys, depending upon whether they were gathered before and after the examinations. These differences, however, applied only to students who had not passed the examinations. It is important to consider whether the period in which the surveys can be completed is scheduled to take place before or after the examinations.

In summary, the literature shows that electronic SET procedures perform as well as traditional paper-and-pencil evaluation forms do, and that they yield similar results. Although electronic surveys obviously offer considerable advantages, their greatest challenge continues to involve increasing the response rate.

Concurrent validity of online ratings of professors. In recent years, the territory of SET has expanded beyond the exclusive domain of institutions to the World Wide Web through such faculty-rating sites as RateMyProfessors.com, PassCollege.com, ProfessorPerformance.com, Ratingsonline.com, and Reviewum.com (Otto, Sanford, & Ross, 2008). The homepage of the most popular site, RateMyProfessors.com, states that, in 2011, the website counted more than 10 million completed rating forms for more than one million teachers in more than 6,500 (Anglo-Saxon) universities and colleges. The rating form consists of five single-item questions concerning the easiness, clarity, and helpfulness of the teacher, as well as the student's level of interest prior to attending class and the use of the textbook during the course. Students are also asked to provide other information, including the title of the course and their own course attendance and grade, and they have the opportunity to add additional detailed comments about the course or the professor. Finally, students are asked to rate the appearance of the teacher involved as "hot" or "not" (although the website suggests that this rating is "just for fun").

The RateMyProfessors.com website is subject to a noncontrolled self-selection bias (since we can assume that only those students who really liked or disliked a teacher will be more likely to register and to share their experiences via such environments), which has consequences for the representativeness, validity, and reliability of the results (for an overview, see Davison & Price, 2009). Data from these websites should therefore be taken with a grain of salt, and they should not be used for summative evaluations. Nevertheless, many students use these ratings as a

source of information about their teachers (Otto et al., 2008). Researchers have recently begun studying the comments and ratings that are available on the RateMyProfessors website in order to learn more about their validity and their relationship to the more traditional forms of SET (as organized at the institutional level). Silva et al. (2008) found that the focus of ratings and comments on the website were very similar to those obtained through traditional evaluations, as they primarily concern teaching characteristics, personality, and global quality. Otto et al. (2008) observed that the online ratings on the RateMyProfessors website reflected student learning, thus possibly constituting a valid measure of teaching quality. In addition, there were no gender differences in the ratings. Besides, ratings on the RateMyProfessors website show statistically significant positive correlations (that exceed .60) with institutionally based SET (Sonntag, Bassett, & Snyder, 2009; Timmerman, 2008). In general, more lenient instructors receive higher overall quality ratings. Stuber, Watson, Carle, and Staggs (2009) observed that, controlling for other predictors, Instructor's Easiness predicted 50% of the variance in the scores on the Overall Quality measure. Timmerman (2008) found similar results and showed that this association can be partially explained by the fact that student's learning is associated with student conceptions of an instructor's perceived easiness.

As identified by Felton, Mitchell, and Stinson (2004), there is a positive correlation between overall ratings and the leniency and sexiness of instructors (correlations were .61 and .30, respectively). Finally, Freng and Webber (2009) find that the "hotness" variable accounted for almost 9% of the variance in SET scores on the RateMyProfessors website. This might strengthen the argument of those who found relationships between physical attractiveness and SET in institutional-based studies (see, e.g., the above mentioned studies by Feely, 2002; Gurung & Vespia, 2007). Still, Freng and Webber's noted that students rate a teacher's hotness on a dichotomous scale rather than a Likert-type scale, thus failing to capture a broader range of variability in attractiveness. The mixed results of these studies and many methodological concerns (self-selection bias, poorly designed questionnaires, the absence of data on the psychometric properties of the instrumentarium) suggest that student evaluations from these websites should be interpreted with great caution.

Discussion

As demonstrated in the previous sections, SET remains a current yet controversial topic in higher education as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical aspects concerning the validity of SET. This article provides an overview of the recent research on the use and the validity of SET. In this final section, we summarize the most important findings of the present study. We relate these findings to the meta-validation framework for SET (Onwuegbuzie et al., 2009) and formulate several suggestions for further research in the field of SET.

Content-Related Validity

Although SET questionnaires can be assumed to have face validity (Onwuegbuzie et al., 2009), recent SET research has revealed differences in the perspectives that

various stakeholders have of good teaching. Such differences threaten both the item validity and the sampling validity of SET instruments, as it is impossible to gather information concerning the extent to which SET instruments provide adequate and complete representations of particular content areas. The renewed call for a common conceptual framework with regard to effective teaching would offer questionnaire architects the opportunity to test their instruments in these areas of validity as well.

Construct-Related Validity

Structural validity. Our review has further shown that many SET instruments have been subjected to thorough validation procedures, although many of these procedures were conducted after the fact. Useful SET instruments are based on both educational theory and the rigorous investigation of their utility and validity (for examples, see, e.g., Marsh et al., 2009; Onwuegbuzie et al., 2007). Nevertheless, many *ad hoc* SET instruments that have never been tested continue to be used for administrative decision-making. When adapting existing instruments to other educational contexts, users are advised to be very cautious of the applicability paradigm (Marsh & Roche, 1997) and to test the validity of the instrument again in the new context. For example, Rindermann and Schofield (2001) demonstrated the validity and reliability of their instrument across six traditional and technical German universities.

It will also be important to test the long-term stability of SET instruments' scores that have been found valid. For example, because the didactic approaches in many institutions have shifted from teacher-centered toward student-centered, it might be quite important to retest existing SET instruments for their utility within these changed contexts—or to determine whether new instruments are needed. In a similar vein, we should consider the evaluation behavior of students when using the same SET instruments for many years. Repeated use might influence their responses in their "umpteenth" evaluation.

Convergent validity. There is no consensus regarding the strength of the correlation between SET and student achievement. This lack has much to do with the measure of learning (i.e., grades, students' perceptions of learning, test outcomes) that was used in the research literature on this topic. Clayson (2009) argued that the more objective the learning is measured, the lower the association between achievement and SET will be. Nevertheless, student achievement should not be measured solely by grades that students make or their perceptions of learning. For example, Clayson (2009) listed five alternatives for increasing the stringency of controls when mapping student learning: using class means (instead of individual means), using common tests in multiple section courses, conducting pretests and posttests, monitoring performance in future classes, and using standardized tests. In this regard as well, agreements are needed in order to determine student achievement (i.e., which measure(s) can be used to investigate the relationships between student achievement and SET scores).

Discriminant validity and divergent validity. The most prominent topic in the SET literature continues to involve the discriminant validity of SET, given the

frequency with which new bias studies are published. Unfortunately, the sometimes-contradictory findings concerning the relationships (or strength of the relationships) between SET and the characteristics of students, courses, and teachers do not promote any conclusive idea of factors that could potentially bias SET scores. This issue is closely related to the number of control variables included in these studies, the way in which these variables are measured, the various research techniques applied, and the characteristics of the samples. It is very difficult to make valuable statements concerning the generalizability of the results (for instance, concerning global effects sizes of such a characteristic on SET scores), as these results are genuinely mixed based on strong and less strong findings on both sides. In addition, recent studies also address the question of whether personal traits and/or halo effects occur in SET, given the possibility that such evaluations could be influenced by psychodynamic aspects that may have consequences for the interpretation of the results.

Outcome validity. Recent research on the outcome validity of SET provides interesting results concerning the attitudes of both teachers and students toward the utility of SET, as well as their actual practices with regard to completing SET forms and the use of their results for the improvement of teaching. In general, students are willing to participate in SET procedures, although they think that teachers and institutions make little or no use of the results. Teachers agree with the use of SET for personnel decisions, as well as to demonstrate the quality of education at institutions, although they make little use of SET in order to improve their teaching. Moreover, responding to SET appears to be more difficult than many stakeholders may assume. It is therefore important for SET to be conducted with great caution and for teachers to count on peers, colleagues, and administrators when interpreting their SET results. Finally, it is important for SET administrators to be trained in both statistics and educational theory, in addition to being well informed about the SET literature. A skilled administrator can remove many of the concerns that teachers have with regard to SET.

The findings concerning SET and the long-term improvement of teaching suggest that such evaluations alone do not lead to better teaching. For this reason, (a) SET should be embedded within a more holistic approach to the evaluation of teaching, in which teachers make a serious effort to reflect upon the improvement of their teaching in a course; (b) teachers should be able to rely on expert consultation concerning their SET scores; and (c) SET should not be the sole means used to map a teacher's teaching (or progress therein).

Generalizability. When reviewing the literature, it becomes clear that most studies in the field suffer from two important limitations that confine their generalizability since, in general, it can be said that these studies were executed in *a particular setting* using *a particular instrument*. First, it is fair to say that most studies were done using nothing more or less than a well-designed (institutional) SET questionnaire, although some standardized questionnaires (such as SEEQ or CEQ) are widely available. Cross-validation procedures in other institutions are needed to demonstrate the generalizability of these institution-based instruments in other settings. Second, the results of many studies are influenced by the SET practice at

the institutions. It is probable that most of the contradictory research results on SET are (at least partly) due to the great variety of methods, measures, controlling variables, SET instruments, and populations used in these studies.

Criterion-Related Validity

SET research reveals a positive correlation between SET scores and other indicators of teaching quality (e.g., student learning outcomes, alumni ratings, self-ratings). This supports the criterion-related validity of scores on SET instruments. Little is known, however, concerning whether the various well-validated SET instruments (e.g., the SEEQ or the CEQ) yield similar results when adopted in identical SET settings. Multitrait–multimethod analysis (in which these instruments are used as different measures of several dimensions of effective teaching) or, more simply, analysis of the correlations between the scores generated by the instruments could yield further evidence on the concurrent validity of these instruments.

Online SET has become the norm at many institutions of higher education. This development has understandably generated many studies on the validity of the results from Web-based student evaluations. For institutions, the results obtained with online SET instruments are similar to those obtained with paper-and-pencil instruments, although students provide more comments in an online environment. Low response rates constitute a major disadvantage of online SET, and this has consequences for the interpretation of the results (e.g., it is not clear whether they are representative of the entire population). It would be interesting to learn (a) which types of students participate in SET and which do not and (b) whether the perceptions of participants differ from those of nonparticipants. Researchers have found that internet-based SET systems yield results that are comparable to those obtained within the institutions. We nevertheless advise against relying on these websites, due to self-selection bias on the part of students, the psychometric properties of the instruments used, and the relationship between SET results and teacher characteristics that are unrelated to effective teaching (e.g., their hotness or sexiness).

Conclusion

This review of the state of the art in the literature has shown that the utility and validity ascribed to SET should continue to be called into question. Next to some, although much-researched, topics such as the dimensionality debate and the bias question, new research lines are delineated (i.e., the utility of online SET, teacher personal characteristics affecting SET). Our systematic use of the meta-validity framework of Onwuegbuzie et al. (2009), however, shows that many types of validity of SET remain at stake. Because conclusive evidence has not been found yet, such evaluations should be considered fragile, as important stakeholders (i.e., the subjects of evaluations and their educational performance) are often judged according to indicators of effective teaching (in some cases, a single indicator), the value of which continues to be contested in the research literature.

References

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. *New Directions for Teaching and Learning*, *43*, 97–111. doi:10.1002/tl.37219904309

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness-generalizability of '*N* = 1' research. Comment on Marsh (1991). *Journal of Educational Psychology*, *83*, 411–415. doi:10.1037/0022-0663.83.3.411

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153–166. doi:10.1023/A:1008168421283

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431–441. doi:10.1037/0022-3514.64.3.431

*Anderson, H. M., Cain, J. C., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, *69*, 34–43. Retrieved from http://archive.ajpe.org/view.asp?art=aj690105&pdf=yes

*Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education*, *30*, 723–748. doi:10.1080/03075070500340101

*Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research*, *48*, 215–224. doi:10.1016/j.ijer.2009.10.001

*Arthur, L. (2009). From performativity to professionalism: Lecturer's responses to student feedback. *Teaching in Higher Education*, *14*, 441–454. doi:10.1080/1356251090305022

*Balam, E., & Shannon, D. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment and Evaluation in Higher Education*, *35*, 209–221. doi:10.1080/02602930902795901

*Ballantyne, C. (2003). Online evaluations of teaching: An examination of current practice and considerations for the future. *New Directions for Teaching and Learning*, *96*, 103–112. doi:10.1002/tl.127

*Ballantyne, R., Borthwick, J., & Packer, J. (2000). Beyond student evaluation of teaching: Identifying and addressing academic staff development needs. *Assessment and Evaluation in Higher Education*, *25*, 221–236. doi:10.1080/713611430

*Barnes, D., Engelland, B., Matherne, C., Martin, W., Orgeron, C., Ring, J., et al. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*, 199-213.

*Barkhi, R., & Williams, P. (2010). The impact of electronic media on faculty evaluation. *Assessment and Evaluation in Higher Education*, *35*, 241–262. doi:10.1080/02602930902795927

*Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business*, *84*, 40–46. doi:10.3200/JOEB.84.1.40-46

*Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, *18*, 91–106. doi:10.1007/s11092-006-9001-8

*Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*, 25–35. doi:10.1111/j.1471-6402.2006.00259.x

*Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*, 253–265. doi:10.1016/j.econedurev.2006.08.007

Beecham, R. (2009). Teaching quality and student satisfaction: Nexus or simulacrum? *London Review of Education*, *7*, 135–146. doi:10.1080/14748460902990336

*Beran, T. N., & Rokosh, J. L. (2009). Instructor's perspectives on the utility of student ratings of instruction. *Instructional Science*, *37*, 171–184. doi:10.1007/s11251-007-9045-2

*Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education*, *30*, 593–601. doi:10.1080/02602930500260688.

*Beran, T., Violato, C., Kline, D., & Frideres, J. (2009). What do students consider useful about student ratings? *Assessment and Evaluation in Higher Education*, *34*, 519–527. doi:10.1080/02602930802082228

*Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. (2004). Interpreting course evaluation results: Insights from thinkaloud interviews with medical students. *Medical Education*, *38*, 1061–1070. doi:10.1111/j.1365-2929.2004.01953.x

Blackmore, J. (2009). Academic pedagogies, quality logics and performative universities: Evaluating teaching and what students want. *Studies in Higher Education*, *34*, 857–872. doi:10.1080/03075070902898664

Bolivar, A. (2000). Student teaching evaluations: Options and concerns. *Journal of Construction Education*, *5*, 20–29. Retrieved from http://www.ascjournal.ascweb.org/

*Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Journal of Economic Education*, *32*, 3–17. doi:10.1080/00220480109595166

*Bothell, T. W., & Henderson, T. (2003). Do online ratings of instruction make sense? *New Directions for Teaching and Learning*, *96*, 69–79. doi:10.1002/tl.124

*Braun, E., & Leidner, B. (2009). Academic course evaluation. Theoretical and empirical distinctions between self-rated gain in competences and satisfaction with teaching behavior. *European Psychologist*, *14*, 297–306. doi:10.1027/1016-9040.14.4.297

*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*, 289–306. doi:10.1007/s11092-011-9126-2

*Bullock, C. D. (2003). Online collection of midterm student feedback. *New Directions for Teaching and Learning*, *96*, 95–102. doi:10.1002/tl.126

*Burden, P. (2008). Does the end of semester evaluation forms represent teacher's views of teaching in a tertiary education context in Japan? *Teaching and Teacher Education*, *24*, 1463–1475. doi:10.1016/j.tate.2007.11.012

*Burden, P. (2010). Creating confusion or creative evaluation? The use of student evaluation of teaching surveys in Japanese tertiary education. *Educational Assessment, Evaluation and Accountability*, *22*, 97–117. doi:10.1007/s11092-010-9093-z

*Burdsal, C. A., & Bardo, J. W. (1986). Measuring student's perception of teaching: Dimensions of evaluation. *Educational and Psychological Measurement*, *46*, 63–79. doi:10.1177/0013164486461006

*Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education*, *33*, 567–576. doi:10.1080/02602930701699049

*Campbell, H., Gerdes, K., & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management*, *24*, 611–620. doi:10.1002/pam.20122

Cashin, W. E., & Perrin, P. B. (1978). *IDEA Technical Report No. 4. Description of IDEA Standard Form Data Base*. Manhattan, KS: Center for Faculty Evaluation and Development in Higher Education.

Centra, J. A. (1998). *Development of The Student Instructional Report II*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Products/283840.pdf

*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. doi:10.1023/A:1025492407752

*Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*, 17–33. Retrieved from www.jstor.org/stable/2649280

*Chen, Y., & Hoshower, L. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education*, *28*, 71–88. doi:10.1080/02602930301683

*Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching. *Structural Equation Modeling*, *7*, 442–460. doi:10.1207/S15328007SEM0703_5

*Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, 16–30. doi:10.1177/0273475308324086

*Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*, 149–160. doi:10.1177/0273475306288402

*Coffey, M., & Gibbs, G. (2001). The evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education*, *26*, 89–93. doi:10.1080/02602930020022318

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281–309. doi:10.3102/00346543051003281

*Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education*, *30*, 123–136. doi:10.1080/02602930420002642 3

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354. doi:10.1037/h0047358

Crumbley, L. C., Flinn, R. E., & Reichelt, K. J. (2010). What is ethical about grade inflation and coursework deflation? *Journal of Academic Ethics, 8*, 187–197. doi:10.1007/s10805-010-9117-9

*Darby, J. A. (2008). Course evaluations: A tendency to respond "favourably" on scales? *Quality Assurance in Education*, *16*, 7–18. doi:10.1108/09684880810848387

*Davidovitch, N., & Soen, D. (2006a). Class attendance and students' evaluation of their college instructors. *College Student Journal*, *40*, 691–703.

*Davidovitch, N., & Soen, D. (2006b). Using students' assessments to improve instructors' quality of teaching. *Journal of Further and Higher Education*, *30*, 351–376. doi:10.1080/03098770600965375

*Davison, E., & Price, J. (2009). How do we rate? An evaluation of online evaluations. *Assessment & Evaluation in Higher Education*, *34*, 51–65. doi:10.1080/02602930801895695

*Dee, K. C. (2007). Student perceptions of high course workloads are not associated with poor student evaluations of instructor performance. *Journal of Engineering Education*, *96*, 69–78. Retrieved from http://www.jee.org/2007/january/6.pdf

*Delucchi, M. (2000). Don't worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology*, *28*, 220–231. Retrieved from www.jstor.org/stable/1318991

*Dolnicar, S., & Grun, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education*, *31*, 160–172. doi:10.1177/0273475309335267

*Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*, 611–623. doi:10.1080/02602930410001689171

Douglas, J., & Douglas, A. (2006). Evaluating teaching quality. *Quality in Higher Education*, *12*, 3–13. doi:10.1080/13538320600685024

*Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education*, *52*, 717–732. doi:10.1007/s11162-011-9214-7

*Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*, 280–303. doi:10.1177/1052562903027003002

Edström, K. (2008). Doing course evaluation as if learning matters most. *Higher Education Research & Development*, *27*, 95–106. doi:10.1080/07294360701805234

Eiszler, C. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*, 483–501. doi:10.1023/A:1015579817194

Ellis, L., Burke, D., Lomire, P., & McCormack, D. (2003). Student grades and average ratings of instructional quality. The need for adjustment. *The Journal of Educational Research*, *97*, 35–40. doi:10.1080/00220670309596626

*Emery, C. R., Kramer, T. R., & Tian, R. (2003). Return to academic standards: A critique of students' evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*, 37–47. doi:10.1108/09684880310462074

*Ewing, V. L., Stukas, A. A., & Sheehan, E. P. (2003). Student prejudice against male and lesbian lecturers. *The Journal of Social Psychology*, *143*, 569–579. doi:10.1080/00224540309598464

*Feeley, T. H. F. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, *51*, 225–236. doi:10.1080/03634520216519

Feldman, K. A. (1997). Identifying exemplary teachers and teaching. Evidence from student ratings. In R. Perry, & J. Smart (Eds.), *Effective teaching in higher education. Research and Practice* (pp. 368-395). New York: Agathon.

*Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment and Evaluation in Higher Education*, *29*, 91–108. doi:10.1080/0260293032000158180

*Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning*, *87*, 85–100. doi:10.1002/tl.10001

*Freng, S., & Webber, D. (2009). Turning up the heat on online teaching evaluations: Does "hotness" matter? *Teaching of Psychology*, *36*, 189–193. doi:10.1080/00986280902959739

*Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*, 353–374. doi:10.1007/s11162-011-9229-0

*Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluations: Mode can matter. *Assessment & Evaluation in Higher Education*, *30*, 581–592. doi:10.1080/02602930500260647

*Ghedin, E., & Aquario, D. (2008). Moving towards multidimensional evaluation of teaching in higher education: A study across four faculties. *Higher Education*, *56*, 583–597. doi:10.1007/s10734-008-9112-x

*Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychology Reports*, *95*, 1023–1030. doi:10.2466/pr0.95.3.1023-1030

*Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education*, *32*, 603–615. doi:10.1080/03075070701573773

*Goldstein, G. S., & Benassi, V. A. (2006). Students' and instructors' beliefs about excellent lecturers and discussion leaders. *Research in Higher Education*, *47*, 685–707. doi:10.1007/s11162-006-9011-x

*Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe, 89*, 44–46.

*Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality—Analysis of relevant factors based on empirical research. *Assessment & Evaluation in Higher Education*, *28*, 229–238. doi:10.1080/0260293032000059595

*Griffin, B. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology*, *26*, 534–552. doi:10.1006/ceps.2000.1075

*Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology*, *29*, 410–425. doi:10.1016/j.cedpsych.2003.11.001

*Guinn, B., & Vincent, V. (2006). The influence of grades on teaching effectiveness ratings at a Hispanic-serving institution. *Journal of Hispanic Higher Education*, *5*, 313–321. doi:10.1177/1538192706291138

*Gump, S. E. (2007). Student evaluation of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly*, *30*, 55–68.

*Gursoy, D., & Umbreit, W. T. (2005). Exploring students' evaluations of teaching effectiveness: What factors are important? *Journal of Hospitality and Tourism Research*, *29*, 91–109. doi:10.1177/1096348004268197

*Gurung, R., & Vespia, K. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology*, *34*, 5–10. doi:10.1080/00986280709336641

*Haladyna, T., & Amrein-Beardsley, A. (2009). Validation of a research-based student survey of instruction in a college of education. *Educational Assessment, Evaluation and Accountability*, *21*, 255–276. doi:10.1007/s11092-008-9065-8

*Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369–376. doi:10.1016/j.econedurev.2004.07.013

*Harnish, R. J., & Bridges, K. R. (2011). Effect of syllabus tone: Students' perceptions of instructor and course. *Social Psychology of Education*, *14*, 319–330. doi:10.1007/s11218-011-9152-4

*Harrison, P., Douglas, D., & Burdsal, C. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*, 311–323. doi:10.1023/B:RIHE.0000019592.78752.da

*Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal*, *40*, 588–596.

*Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology*, *93*, 790–796. doi:10.1037/0022-0663.93.4.790

*Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, *36*, 29–42. doi:10.3200/JECE.36.1.29-42

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of student's perceptions of teaching effectiveness. *Educational and Psychological Measurement*, *59*, 580–596. doi:10.1177/00131649921970035

Jauhiainen, A., Jauhiainen, A., & Laiho, A. (2009). The dilemmas of the "efficiency university" policy and the everyday life of university teachers. *Teaching in Higher Education*, *14*, 417–428. doi:10.1080/13562510903050186

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, *5*, 419–434. doi:10.1080/713699176

*Johnson, T. D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning*, *96*, 49–59. doi:10.1002/tl.122

*Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology*, *37*, 16–20. doi:10.1080/00986280903426282

*Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, *33*, 84–90. doi:10.1207/s15328023top3302_1

*Kember, D., Jenkins, W., & Kwok, C.N. (2004). Adult students' perceptions of good teaching as a function of their conceptions of learning—Part 2. Implications for the evaluation of teaching. *Studies in Continuing Education*, *26*, 81–97. doi:10.1080/158037042000199461

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

*Kember, D., & Leung, D. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education*, *52*, 279–299. doi:10.1007/s11162-010-9194-z

*Kember, D., Leung, D., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education*, *27*, 411–425. doi:10.1080/0260293022000009294

*Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education*, *40*, 69–97. doi:10.1023/A:1004068500314

*Kim, C., Damewood, E., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education*, *24*, 458–473. doi:10.1177/105256290002400405

Knapper, C. (2001). Broadening our approach to teaching evaluation. *New Directions for Teaching and Learning*, *88*, 3–9. doi:10.1002/tl.32

*Kogan, L., Schoenfeld-Tacher, R., & Helleyer, P. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*, 623–636. doi:10.1080/13562517.2010.491911

*Kohn, J., & Hartfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal*, *10*, 121–137.

*Kreiter, C. D., & Laksham, V. (2005). Investigating the use of sampling for maximising the efficiency of student-generated faculty teaching evaluations. *Medical Education*, *39*, 171–175. doi:10.1111/j.1365-2929.2004.02066.x

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research*, *27*, 9–25. doi:10.1002/ir.1

*Landrum, R. E., & Braitman, K. A. (2008). The effect of decreasing response options on students' evaluation of instruction. *College Teaching*, *56*, 215–217. doi:10.3200/CTCH.56.4.215-218

*Lang, J. W. B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*, 187–205. doi:10.1007/s11251-006-9006-1

*Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*, 417–428. doi:10.1016/j.econedurev.2006.12.003

Larsen, M. A. (2005). A critical analysis of teacher evaluation policy trends. *Australian Journal of Education*, *49*, 292–305.

Lattuca, L., & Domagal-Goldman, J. (2007). Using qualitative methods to assess teaching effectiveness. *New Directions for Institutional Research*, *136*, 81–93. doi:0.1002/ir.233

*Layne, B. H., Decristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*, 221–232. doi:10.1023/A:1018738731032

*Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education*, *46*, 571–591. doi:10.1007/s11162-005-3365-3

*Liu, Y. (2006). A comparison study of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology and Distance Learning*, *4*, 15–29. Retrieved from http://www.itdl.org/

*Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, *22*, 108–119. doi:10.1177/0273475300222005

Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77–95. doi:10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, *76*, 707–754. doi:10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research. *International Journal of Educational Research*, *11*, 253–388. doi:10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami & d'Apollonia (1991). *Journal of Educational Psychology*, *83*, 416–421. doi:10.1037/0022-0663.83.3.416

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285–296. doi:10.1037/0022-0663.83.2.285

*Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183–212. doi:10.3102/00028312038001183

*Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology*, *99*, 775–790. doi:10.1037/0022-0663.99.4.775

*Marsh, H. W. (2007b). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York: Springer.

Marsh, H. W., & Hovecar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education*, *7*, 9–18. doi:10.1016/0742-051X(91)90054-S

*Marsh, H. W., Muthèn, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. doi:10.1080/10705510903008220

Marsh, H. W., & Roche, L A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, *52*, 1187–1197. doi:10.1037/0003-066X.52.11.1187

*Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology*, *92*, 202–228. doi:10.1037/0022-0663.92.1.202

*Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology*, *33*, 176–179. doi:10.1207/s15328023top3303_4

McKone, K. E. (1999). Analysis of student feedback improves instructor effectiveness. *Journal of Management Education*, *23*, 396–415. doi:10.1177/105256299902300406

*McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education*, *37*, 3–20. doi:10.3200/JECE.37.1.3-20

*McPherson, M. A., & Todd Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*, 868–881. doi:10.1111/j.1540-6237.2007.00487.x

*McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*, 37–51. doi:10.1057/palgrave.eej.9050042

*Menges, R. J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. *New Directions for Teaching and Learning*, *83*, 5–11. doi:10.1002/tl.8301

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

*Mohanty, G., Gretes, J., Flowers, C., Algozzine, B., & Spooner, F. (2005). Multimethod evaluation of instruction in engineering classes. *Journal of Personnel Evaluation in Higher Education*, *18*, 139–151. doi:10.1007/s11092-006-9006-3

Molesworth, M., Nixon, E., & Scullion, R. (2009). Having, being and higher education: The marketisation of the university and the transformation of the student into consumer. *Teaching in Higher Education*, *14*, 277–287. doi:10.1080/13562510902898841

*Montoneri, B., Lee, C. C., Lin, T. T., & Huang, S. L. (2011). A learning performance evaluation with benchmarking concept for English writing courses. *Expert Systems with Applications*, *38*, 14542–14549. doi:10.1016/j.eswa.2011.05.029

*Montoneri, B., Lin, T. T., Lee, C. C., & Huang, S. L. (2012). Application of data envelopment analysis on the indicators contributing to learning and teaching performance. *Teaching and Teacher Education*, *28*, 382–395. doi:10.1016/j.tate.2011.11.006

*Moore, S., & Kuol, N. (2005a). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education*, *10*, 57–73. doi:10.1080/1356251052000305534

*Moore, S., & Kuol, N. (2005b). A punitive bureaucratic tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMullin (Eds.), *Emerging issues in the practice of university learning and teaching. Part 3: Developing and growing as a university teacher* (pp. 141–146). Dublin, Ireland: University of Limerick.

*Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies*, *35*, 547–552. doi:10.1080/03055690902880299

*Nasser, F., & Fresko, B. (2001). Interpreting student ratings: Consultation, instructional modification, and attitudes towards course evaluation. *Studies in Educational Evaluation*, *27*, 291–305. doi:10.1016/S0191-491X(01)00031-1

*Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*, 187–198. doi:10.1080/02602930220128751

*Nasser, F., & Fresko, B. (2006). Predicting student ratings: The relationship between actual student ratings and instructor's predictions. *Assessment & Evaluation in Higher Education*, *31*, 1–18. doi:10.1080/02602930500262338

*Nasser, F., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, *35*, 37–44. doi:10.1016/j.stueduc.2009.01.002

*Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*, 301–314. doi:10.1080/02602930701293231

*Ogier, J. (2005). Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education*, *30*, 477–488. doi:10.1080/02602930500186941

Oleinik, A. (2009). Does education corrupt? Theories of grade inflation. *Educational Research Review*, *4*, 156–164. doi:10.1016/j.edurev.2009.03.001

*Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology*, *26*, 382–399. doi:10.1006/ceps.2000.1070

Olivares, O. J. (2003). A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education*, *8*, 233–245. doi:10.1080/1356251032000052465

*Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, *43*, 197–209. doi:10.1007/s11135-007-9112-4

Onwuegbuzie, A. J., & Weems, G. H. (2004). Response categories on rating scales: Characteristics of item respondents who frequently utilize midpoint. *Research in the Schools*, *9*, 73-90.

*Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*, 113–160. doi:10.3102/0002831206298169

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning*, *87*, 3–15. doi:10.1002/tl.23

*Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *109*, 27–44. doi:10.1002/ir.2

*Otto, J., Sanford, D. A., & Ross, D. N. (2008). Does RateMyProfessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*, 355–368. doi:10.1080/02602930701293405

*Pan, D., Tan, G. S. H., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. K. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education*, *50*, 73–100. doi:10.1007/s11162-008-9109-4

*Paswan, A. K., & Young, J. A. (2002). Student evaluation of instructor: A nomological investigation using structural equation modelling. *Journal of Marketing Education*, *24*, 193–202. doi:10.1177/0273475302238042

*Patrick, C. L. (2011). Student evaluations of teaching: Effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, *36*, 239–249. doi:10.1080/02602930903308258

Paulsen, M. B. (2002). Evaluating teaching performance. *New Directions for Institutional Research*, *114*, 5–18. doi:10.1002/ir.42

*Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8*, 399–411. doi:10.1080/13562510309396

*Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, *74*, 215–253. doi:10.3102/00346543074002215

Platt, M. (1993). What student evaluations teach. *Perspectives on Political Science*, *22*, 29–40. doi:10.1080/10457097.1993.9944516

*Pozo-Munoz, C., Rebolloso-Pacheco, E., & Fernandez-Ramirez, B. (2000). The "Ideal Teacher". Implications for student evaluations of teaching effectiveness. *Assessment & Evaluation in Higher Education*, *25*, 253–263. doi:10.1080/02602930050135121

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

Redding, R. (1998). Students' evaluations of teaching fuel grade inflation. *American Psychologist*, *53*, 1227–1228. doi:10.1037/0003-066X.53.11.1227

*Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, *34*, 91–115. doi:10.1080/01411920701492043

Remmers, H., & Brandenburg, G. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519–527.

*Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education*, *46*, 929–953. doi:10.1007/s11162-005-6934-6

*Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, *30*, 387–415. doi:10.1080/02602930500099193

*Richardson, J. T. E. (2012). The role of response biases in the relationship between students' perceptions of their courses and their approaches to studying in higher education. *British Educational Research Journal*, *38*, 399–418. doi:10.1080/01411926.2010.548857

*Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university teaching across courses and teachers. *Research in Higher Education*, *42*, 377–399. doi:10.1023/A:1011050724796

*Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology*, *133*, 19–35. doi:10.3200/GENP.133.1.19-35

*Robertson, S. I. (2004). Student perceptions of student perception of module questionnaires: Questionnaire completion as problem solving. *Assessment and Evaluation in Higher Education*, *29*, 663–679. doi:10.1080/0260293042000227218

*Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science*, *28*, 439–468. doi:10.1023/A:1026576404113

*Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negative worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*, 117–134. doi:10.1080/02602930802618344

*Santhanam, E., & Hicks, O. (2001). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7*, 17–31. doi:10.1080/13562510120100364

*Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*, 401–415. doi:10.1016/j.learninstruc.2006.09.002

Seldin, P. (1993). The use and abuse of student ratings of professors. *Chronicle of Higher Education*, *39*, A40.

*Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*, 397–405. doi:10.1080/713611436

*Silva, K. M., Silva, F. J., Quinn, M. A., Draper, J. N., Cover, K. R., & Munoff, A .A. (2008). Rate my Professor: Online evaluations of psychology instructors. *Teaching of Psychology*, *35*, 71–80. doi:10.1080/00986280801978434

*Simpson, P., & Siguaw, J. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22*, 199–213. doi:10.1177/0273475300223004

*Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication*, *30*, 64–77. doi:10.1080/07491409.2007.10162505

*Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, *34*, 499–504. doi:10.1080/02602930802079463

*Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, *27*, 397–409. doi:10.1080/0260293022000009285

*Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation*, *36*, 121–131. doi:10.1016/j.stueduc.2011.02.001

*Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education. Development of an instrument based on 10 Likert scales. *Assessment and Evaluation in Higher Education*, *32*, 667–679. doi:10.1080/02602930601117191

*Spooren, P., Mortelmans, D., & Thijssen, P. (2012). Content vs. style. Acquiescence in student evaluations of teaching? *British Educational Research Journal*, *38*, 3–21. doi:10.1080/01411926.2010.523453

*Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*, 50.

*Stack, S. (2003). Research productivity and student evaluation of teaching in social science classes. *Research in Higher Education*, *44*, 539–556. doi:10.1023/A:1025439224590

*Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education*, *25*, 269–291. doi:10.1177/105256290102500302

*Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education*, *32*, 403–415. doi:10.1080/02602930600898536

*Stodnick, M., & Rogers, P. (2008). Using SERVQUAL to measure the quality of the classroom experience. *Decisions Sciences Journal of Innovative Education*, *6*, 115–133. doi:10.1111/j.1540-4609.2007.00162.x

*Stuber, J. M., Watson, A., Carle, A., & Staggs, K. (2009). Gender expectations and on-line evaluations of teaching: Evidence from RateMyProfessors.com. *Teaching in Higher Education*, *14*, 387–399. doi:10.1080/13562510903050137

*Timmerman, T. (2008). On the Validity of RateMyProfessors.com. *Journal of Education for Business*, *84*, 55–61. doi:10.3200/JOEB.84.1.55-61

*Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education*, *41*, 637–661. doi:10.1023/A:1007075516271

Titus, J. (2008). Student ratings in a consumerist academy: Leveraging pedagogical control and authority. *Sociological Perspectives*, *51*, 397–422. doi:10.1525/sop.2008.51.2.397

*Toland, M., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272–296. doi:10.1177/001316440426866

*Tom, G., Tong, S. T., & Hesse, C. (2010). Thick slice and thin slice teaching evaluations. *Social Psychology of Education*, *13*, 129–136. doi:10.1007/s11218-009-9101-7

*Tucker, B., Jones, S., Straker, L., & Cole, J. (2003). Course evaluation on the web: Facilitating student and teacher reflection to improve learning. *New Directions for Teaching and Learning*, *96*, 81–94. doi:10.1002/tl.125

Valsan, C., & Sproule, R. (2005). The invisible hands behind the student evaluation of teaching: The rise of the new managerial elite in the governance of higher education. *Journal of Economic Issues*, *42*, 939–958.

*Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, *35*, 101–115. doi:10.1080/02602930802618336

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education*, *23*, 191–210. doi:10.1080/0260293980230207

Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, *34*, 166–176.

*Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology*, *30*, 190–206. doi:10.1016/j.cedpsych.2004.07.003

*Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education*, *26*, 17–30. doi:10.1177/0273475303258276

*Wolfer, T., & Johnson, M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education*, *39*, 111–121.

*Yao, Y., & Grady, M. (2005). How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education*, *18*, 107–126. doi:10.1007/s11092-006-9000-9

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education*, *12*, 55–76. doi:10.1080/13562510601102131

Authors

PIETER SPOOREN holds master's degrees in Educational Sciences and Quantative Analysis in the Social Sciences and a PhD in Social Sciences. He is affiliated as an educational advisor at the Faculty of Political and Social Sciences of the University of Antwerp (Belgium). His particular activities are educational innovation and evaluation of the educational process and of educators. His main research interests focus on students' evaluation of teaching (SET), in particular their use and validity.

BERT BROCKX holds a master's degree in Educational Sciences. He is affiliated as a predoctoral researcher at the Faculty of Political and Social Sciences of the University of Antwerp (Belgium). His main research interests focus on the validity of students' evaluation of teaching (SET).

DIMITRI MORTELMANS is an associate professor at the University of Antwerp. He is head of the Research Center for Longitudinal and Life Course Studies (CELLO). He publishes in the domain of family sociology and sociology of labor. Important topics of his expertise are ageing, divorce, and gender differences in career trajectories.