Assessment & Evaluation in Higher Education
Vol. 37, No. 6, September 2012, 683–699

Routledge
Taylor & Francis Group

Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use

Stephen L. Wright* and Michael A. Jenkins-Guarnieri

Department of Counseling Psychology, University of Northern Colorado, 501 20th St, 248 McKee Hall – Box 131, Greeley, CO 80639, USA

There is a plethora of research on student evaluations of teaching (SETs) regarding their validity, susceptibility to bias, practical use and effective implementation. Given that there is not one study summarising all these domains of research, a comprehensive overview of SETs was conducted by combining all prior meta-analyses related to SETs. Eleven meta-analyses were identified, and nine meta-analyses covering 193 studies were included in the analysis, which yielded a small-to-medium overall weighted mean effect size (r = .26) between SETs and the variables studied. Findings suggest that SETs appear to be valid, have practical use that is largely free from gender bias and are most effective when implemented with consultation strategies. Research, teaching and policy implications are discussed.

Keywords: student evaluations of teaching; student achievement; instructor feedback

Given the near-ubiquitous use of student evaluations of teaching (SETs) in American institutions of higher education (d'Apollonia and Abrami 1997; Dommeyer et al. 2002; Richardson 2005; Shevlin et al. 2000;), it is no surprise that there have been much research and debate on SETs. In a commonly cited example, Seldin (1993) reported that 86% of higher education personnel decisions use SETs as a major criterion, representing a usage increase of 18% since 1984 (Seldin 1984). Faculty and teaching evaluation by students has sparked more than 80 years of investigation (Clayson 2009), yielding well over 2000 relevant published articles listed in popular online databases (Centra 2003). This confirms that SETs are an important area of research and that more well-wrought investigations are needed to clarify the intense debates (e.g. Greenwald and Gillmore 1997; Griffin 2004; Marsh 1987; Marsh and Roche 2000) over their validity, susceptibility to biasing factors, practical use and effective implementation. Given a tremendous number of SET studies, multiple meta-analyses have attempted to synthesise these areas of research; however, there is not a single outlet that combines and summarises all the meta-analyses. This study attempts to address the aforementioned issues by providing a comprehensive overview of SETs by combining all prior meta-analyses related to SETs. Particularly, meta-analyses were identified and examined to ascertain the variables related to SETs (i.e. student achievement, gender of instructor, feedback/consultation) and the overall impact of SET ratings with all variables investigated.

*Corresponding author. Email: stephen.wright@unco.edu

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

http://dx.doi.org/10.1080/02602938.2011.563279
http://www.tandfonline.com

Construct validity

Although most researchers seem to accept the need for some form of instructor evaluation in college courses, much debate centers on the validity of SETs. Clayson (2009, 17) noted that these evaluations are based on the idea that 'students will learn more from good teachers', using this association between learning and teaching effectiveness as justification for SET use by implying their theoretical construct-validity. Cohen (1981) noted that the complex factors involved in the research on SETs make it difficult to reach decisive conclusions or strong generalisations about SET validity. However, he states that researchers in this field generally consider student-learning to be the most significant gauge for measuring teaching effectiveness. In his extensive review, Marsh (1987) supported this idea in finding consistent and significant relationships between student achievement and student ratings of instructors. After nearly 30 years of work on SETs, Marsh (2007b, 339) still upholds the validity of SETs because 'the sections that evaluate the teaching as most effective are also the sections that perform best on standardised final examinations'. Meta-analyses and large-scale reviews have both supported and questioned the validity of SETs. Cohen (1981, 305) found that; 'student ratings of instruction are a valid index of instructional effectiveness', while d'Apollonia and Abrami (1997) report that multisection validity studies support the use of SETs as well as their validity. However, Onwuegbuzie, Daniel, and Collins (2009, 206) concluded with 'serious doubt on the score validity of STEs [*sic*], particularly in terms of areas of content-related validity and construct-related validity', and Pounder (2007) echoed these concerns over the utility and validity of SETs.

Potential biasing variables

Much of the recent research questioning the use of SETs centers around potential biasing variables that could unduly impact SET ratings. One of the most prominent areas of discussion focuses on the potential for expected grades and an instructor's grading practices to influence student ratings of both instruction and course quality (for an extensive discussion of research on the relationship between grades and SETs, see Gump 2007). Authors such as Greenwald and Gillmore (1997) promote a grading leniency hypothesis, although this effort may suffer from experimenter bias (Gump 2007) and methodological issues (Marsh and Roche 2000). These authors argue that instructors can earn higher SET ratings by following a more lenient grading policy or 'buying' better ratings by giving higher grades. Recent research supports this hypothesis as well (Griffin 2004). In contrast, other researchers (Centra 2003; Marsh and Roche 2000) promote a validity hypothesis, theorising that student ratings reflect student perceptions of learning and are not significantly influenced by expected course grade. This hypothesis has been supported by more recent work of sound design and quality (Remedios and Lieberman 2008). Current research puts this debate into perspective, as the relationship between student ratings and grades is small (Boysen 2008), with estimates at about a .20 correlation (Marsh and Roche 2000). Although, based on Cohen's (1988) correlational effect size guidelines, a .20 correlation falls between a small (.1)-to-moderate (.3) effect. Subsequently, in a study with a sample that included approximately 55,000 classes among two-year colleges and four-year colleges, Centra (2003) found a small correlation of .11 between expected grades and student ratings. Indeed, Marsh (1987) concluded that the impacts of potential biasing factors on SETs (e.g. expected grade, course workload) were relatively weak. Given the level of contradiction and ambiguity in the research on expected grades and grading

practices as biasing variables affecting SETs (Gump 2007), perhaps the effects of these variables are quite small, inconsistent or even confounded by other variables. In summarising a thorough review of research relevant to the grade/SET relationship, Marsh concludes that 'evidence from a variety of different studies clearly supports the validity and student characteristics hypotheses. Whereas a grading-leniency effect may produce *some* bias in SETs, support for this suggestion is weak, and the size of such an effect is likely to be insubstantial' (2007b, 357, emphasis in original).

Marsh (2007b) suggests that a significant number of studies addressing the impact of bias on SETs suffer from methodological flaws and poor operationalisation of bias among other limitations (c.f. Marsh 1987; Marsh and Dunkin 1997). In this chapter, he summarises the potential biasing variables considered in the SET literature and suggests that the relationships between SET scores and potential biasing variables are relatively small, with only a limited number of studies finding a correlation of .30 or above, whereas the typical relationships were substantially lower (Marsh and Roche 1997). Furthermore, Marsh and Dunkin (1997) found that prior subject interest was the variable most strongly correlated with SETs, acknowledging that this variable impacts some elements of effective teaching as well. Addressing another common faculty concern over SETs, course workload was found to be *positively* (italics added) correlated with SET scores (Marsh and Roche 2000), although more recent literature suggests that this relationship is more complex and perhaps curvilinear (Centra 2003). However, Marsh and Roche report that even potentially biasing perceptions of expected grades and workload difficulty remained relatively stable over time, as did correlations between overall ratings, expected grades and workload.

Additional work on biasing variables on SETs has focused on instructor gender or sex, attractiveness, and nonverbal behaviour, among other factors. For example, recent studies on gender bias reported that female students rate instructors of the same sex significantly higher on SETs (Bachen, McLoughlin, and Garcia 1999; Centra and Gaubatz 2000). Physical attractiveness has been shown to influence SET ratings as well (Riniolo et al. 2006), although the magnitude of this biasing variable's impact on SETs is still debated (Campbell, Gerdes, and Steiner 2005). In an interesting study in which female undergraduates rated brief videos of instructors lecturing (audio muted), Ambady and Rosenthal (1993) found that teacher ratings by students observing only short clips of nonverbal behaviours were closely associated with teacher ratings by students who had extended interactions with the same instructors. Although a myriad of other factors may have the potential to moderate the relationship between learning and SETs, the present study focused on synthesising all published meta-analyses relevant to SET research.

Consultation strategies

An important caveat to the support of SETs is the question that how they are implemented and used in practice. Many institutions require faculty to use SETs as part of the evaluation process towards their promotion and tenure. Although commonly used for administrative purposes, SETs can be used in also other formative ways, especially in improving quality of teaching. It appears that this is best done by providing feedback to course instructors based on the course SET results paired with some forms of consultation (Marsh 2007b), such as advisory or educational consultation (Penny and Coe 2004). Subsequently, these forms of consultation may include extended discussions of teaching effectiveness (average of two hours), training or workshops on

teaching issues and using multiple sources of data collection when discussing teaching effectiveness (e.g. classroom observations, videotapes of teaching, interviews; Penny and Coe 2004). Furthermore, academic peers who are held in high regard for exceptional teaching are most likely best suited to serve as consultants (Penny and Coe 2004), although just the use of consultation with SETs is beneficial. Menges and Brinko (1986) conducted a meta-analysis and presented a paper on student ratings and consultation. They found a larger effect size on future student ratings in studies combining consultation with student rating feedback ($d = 1.10$), with .80 being a large effect size for group comparisons (Cohen 1988). Previous meta-analyses echo similar findings as well (e.g. Cohen 1980).

Purpose of the study

To date, the authors were unable to find a detailed and thorough examination of all meta-analyses that have been conducted in the area of SETs. Therefore, we designed the current study to provide a comprehensive overview of all meta-analyses related to SETs, which will enable educators, researchers and policy-makers to examine a wealth of research and understand SET validity, its practical use and effective implementation synthesised from critical investigation on the topics of interest. Specifically, each meta-analysis was examined to identify the construct of interest related to SETs (i.e. student achievement as measured by final exams or final grades, feedback/consultation, gender of instructor) and analysed to determine the overall relationship between SET ratings and constructs across all meta-analyses with similar focuses. Given the enormous amount of research on SETs, the purpose of this study is to provide a single outlet that summarises all prior meta-analyses related to SETs.

Method

Identification of studies

A thorough review of the literature was conducted to identify all published meta-analyses that address some aspect of student evaluation of teaching. Literature search procedures included the following databases: ERIC, PsycINFO and Sociological Abstracts. In order for findings to be most applicable, searches were restricted by publication date (from 1975 to present) and by language (English language journals only). The authors followed a systematic and reproducible literature search protocol, the specific steps and procedures of which were well documented in detail and are available upon request. First, the appropriate database descriptors were identified using each database's thesaurus. The authors used the following search terms (asterisks included to account for variations in spelling and pluralisation): 'student evaluation',*[1] 'teaching* performance', 'student evaluation of teaching', 'teacher effectiveness', 'teacher effectiveness evaluation', 'teacher evaluation', 'personnel evaluation', 'literature review*' and 'meta analysis'. Second, each article database was searched using the aforementioned terms in various combinations as keywords, descriptors, title elements and abstract elements while attempting to limit the results to articles defined as a meta-analysis or a review in each database record's title, descriptor or abstract sections. Lastly, the reference sections of relevant literature reviews, meta-analyses and seminal articles on SETs were searched for appropriate meta-analyses.

Criteria for inclusion of studies

The criteria for being included in this literature review of meta-analyses were as follows: (1) the meta-analysis must have been published between 1975 and present, (2) only peer-reviewed, published meta-analysis articles were included, (3) the meta-analysis must have reported an effect size, and (4) the meta-analysis had to include some form of college students' evaluation of teachers. Studies of the following types were excluded: newsletters, conference presentations, unpublished papers, literature reviews and books.

Results

Based on the criteria for inclusion in the study, the authors identified 11 meta-analyses related to student evaluation of teaching (see Table 1). Combining all the meta-analyses, the total number of studies was 233 and included 1330 courses/sections (Clayson 2009, had 1115 sections). The mean number of studies per meta-analysis was 21.18 (SD = 12.01). When combining all of the studies included in the meta-analyses, the earliest publication was in 1936 and the most recent in 2007; the median year was 1976, the mean year was 1974 (SD = 9.45). The following variables were examined among the 11 meta-analyses: student evaluation of teaching (n = 11), student achievement (n = 6), feedback/consultation strategies (n = 3), course ratings (n = 2), instructor expression (n = 1), instructor gender (n = 1) and lecture content (n = 1).

Meta-analyses can best be understood by their effect sizes. Effect sizes are standardised numbers used to capture the amount of difference between two groups or the effect that one variable has on another variable. To help interpret effect sizes, Cohen (1988) provided guidelines for mean difference effects and correlation effects. Cohen's guidelines for correlation effect sizes are as follows: r = .10 is a small effect; r = .30 is a medium effect; and r = .50 or above indicates a large effect. Accordingly, a large effect (r = .50) indicates that 25% of the variance in one variable is accounted for by the other variable of interest. Cohen's (1988) guidelines for the standardised mean difference between two groups specify the amount of difference in magnitude between groups, which are small (d = .20), medium (d = .50) and large (d = .80). Based on these guidelines, a large effect (d = .80) can be interpreted by stating that the mean of one group is at the 79th percentile relative to the mean of the comparison group; an effect size of .00 indicates that the mean of a group is at the 50th percentile relative to the mean of the comparison group (Cohen 1988).

In a meta-analysis with 41 studies, Cohen (1981) found a medium-to-large correlation effect size (r = .47), according to Cohen's (1988) guidelines, between overall course rating and student achievement. This finding indicates that 22% of the variance in overall course rating is accounted for by the variance in student achievement. Additionally, overall instructor rating had a medium effect (r = .43) on student achievement (Cohen 1981). As a follow up to his 1981 study, Cohen (1982) further demonstrated a medium-to-large (r = .44) relationship between instructor ratings and student achievement. He then concluded that students who score high on achievement tend to give their teachers higher instructional ratings (1982). Cohen (1983) again confirmed the evidence between student ratings of instruction and student achievement by finding a medium effect (r = .38). Dowell and Neal (1982) combined six studies to investigate the relationship between student achievement and student instructor ratings. They concluded that student achievement had a small-to-medium effect (r = .20) on

Table 1. Summary of meta-analyses included.

Author(s)	Year	Meta-analysis topic and student evaluations of teaching	Number of studies[a]	Range of years[b]	Effect size	Summary
Abrami, Leventhal, and Perry	1982	Instructor expressiveness and lecture content	12	1975–1982 (1979)	Instructor expressiveness effects on student ratings (ω^2 = .293) and on student achievement (ω^2 = .043) Lecture content effects on student achievement (ω^2 = .155) and student ratings (ω^2 = .038) *weighted mean ω^2 reported	Instructor expressiveness has a large effect on student ratings and a small effect on student achievement. Lecture content has a large effect on student achievement and a small effect on student ratings.
Clayson	2009	Student achievement	42 (1115)	1953–2007 (1978)	Weighted mean r = .134 (not significant at .05 level)	Used a voting method to demonstrate a small relationship exists between learning (i.e. test results) and student evaluations of teaching, as 13 of 42 studies found significant positive correlations.
Cohen	1980	Feedback/Consultation	17 (22)	1969–1979 (1977)	d_+ = .38 end of semester SETs improved if SETs were provided at midterm r = .19; calculated using Cohen's formula[c]	SETs at midterm had a small-to-medium effect on SETs at the end of the semester.
Cohen	1981	Student achievement	41 (68)	1949–1979 (1974)	Mean r = .47 (95% CI = .09–.73) overall course rating and student achievement – based on 22 courses Mean r = .43 (95% CI = .21–.61) overall instructor rating and student achievement – based on 67 courses	Results demonstrate the validity and support for using student ratings to measure instructor and course effectiveness.

Table 1. (*Continued*).

Author(s)	Year	Meta-analysis topic and student evaluations of teaching	Number of studies[a]	Range of years[b]	Effect size	Summary
Cohen	1982	Student achievement	16 (21)	1952–1979 (1975)	Mean $r = .44$ (95% CI = .24–.60) for overall instructor rating and student achievement; based on 21 courses	Students who score high on achievement give their teachers higher instructional ratings.
Cohen	1983	Student achievement	18 (33)	1949–1979 (1975)	Mean $r = .38$ (95% CI = .31–.45) student rating of instruction and student achievement	A medium effect was found between student ratings of instruction and student achievement.
Dowell and Neal	1982	Student achievement	6	1950–1975 (1973)	Mean $r = .20$ student achievement and student instructor ratings	Student achievement had a small-to-medium effect on student instructor ratings.
Feldman	1993	Gender of instructor	28	1936–1990 (1977)	Mean $r = .022$	Female instructors were provided slightly higher overall student evaluations.
L'Hommedieu, Menges, and Brinko	1990	Feedback/Consultation	28	1971–1985 (1978)	Written (16 studies), Glass's delta range, $\Delta = -.107–.545$; Personal (six studies), $\Delta = -.10–.50$; Consult (six studies), $\Delta = .00–1.63$) No overall effect size was reported	Small positive effect on teacher ratings when provided with written feedback. Considerable increased effect on teacher ratings when personal consultation is also included in the feedback.
McCallum	1984	Student achievement	14	1953–1980 (1976)	Weighted mean $r = .318$ between instructor rating and student performance/achievement (14 studies) Weighted mean $r = .252$ between course rating and student performance/achievement (four studies)	Medium effect between student instructor rating and student performance. A small-to-medium effect between course rating and student performance/achievement.

Table 1. (*Continued*).

Author(s)	Year	Meta-analysis topic and student evaluations of teaching	Number of studies[a]	Range of years[b]	Effect size	Summary
Penny and Coe	2004	Feedback/Consultation	11	1976–2001 (1980)	d_+ = .69 (95% CI = .43–.95) r = .33; calculated using Cohen's formula[c]	Teachers who were provided with consultation/feedback over their student ratings at midterm largely improved their teaching effectiveness.

[a]Sections listed in parentheses;

[b]Median year listed in parentheses; and

[c]Cohen's (1988, 23) formula, $r = d/[(d^2 + 4)^{.5}]$, was used to convert Cohen's d effect size to Pearson's r effect size.

Note: CI = Confidence intervals and were reported when provided in the meta-analysis.

student instructor ratings. McCallum's (1984) meta-analysis of 14 studies examined how students' ratings of their instructor and overall course rating were related to students' class performance. He found a medium effect between student instructor rating (r = .32) and student performance and a small-to-medium effect (r =.25) between course rating and student performance. Thus, 10.1% of the variance in instructor rating and 6.4% of the variance in course rating can be accounted for by student performance.

Cohen (1988) also established guidelines for omega-squared, which are: ω^2 = .010 for a small effect, ω^2 = .059 for a medium effect and ω^2 = .138 for a large effect. Abrami, Leventhal, and Perry (1982) used weighted mean omega-squared effects sizes to account for the number of students per study. They found that instructor expressiveness explained a large proportion of the variance in student ratings (ω^2 = .29), but a small-to-medium proportion of the variance in student achievement (ω^2 = .04). Abrami, Leventhal, and Perry also reported that lecture content had a large effect on student achievement (ω^2 = .16), but a small-to-medium effect on student ratings (ω^2 = .04). In reviewing 42 studies, Clayson (2009) suggested that there is a small relationship between learning (i.e. test results) and SETs. Although the relationship was not significant at the .05 level, Clayson claimed that a small relationship existed because 13 of the 42 studies found a significant positive correlation, and one study found a negative correlation. It should be noted that Clayson's meta-analysis was the only study that did not find a significant effect size.

In a meta-analysis conducted to combine SET feedback with personal consultation, Cohen (1980) found that students' ratings of instructors at midterm had a small-to-medium effect (d_+ = .38) on students' ratings of instructors at the end of the semester. L'Hommedieu, Menges, and Brinko (1990) reported that written feedback for instructors combined with personal consultation had a far greater effect than written feedback alone. They based their conclusions on 16 studies that provided written feedback (Glass's delta range, Δ = –.107 – .545), and six studies that used consultation in addition to written feedback (Δ = .00 – 1.63). More recently, Penny and Coe (2004) published a meta-analysis using 12 studies investigating the potential impact of consultation strategies on teaching effectiveness and found a moderate effect size (d_+ = .69) among the 11 studies included; one study was excluded due to its artificially deflating the overall mean effect. Feldman (1993) reviewed 28 studies and found that instructor gender slightly impacted students' evaluations of their instructors (r = .02). Although, nine of the 28 studies used each individual student in the class as a unit of analysis, as opposed to the 19 studies that used the average rating of the instructor by all the students in the class. Considering this, the effect size remained virtually meaningless (r = .03) when including only 19 studies.

In the present study, we extensively reviewed each meta-analysis related to SETs; next we calculated new effect sizes by combining the related SET meta-analyses previously reviewed. Given that student evaluation of teaching was the topic included in all meta-analyses, we were interested in examining the variables included in the meta-analyses that impacted student evaluation of teaching (see Table 1). Cohen's d was reported in two studies, omega-squared and Glass's delta were reported in one study each, and the remaining seven studies reported Pearson's r values for effect sizes. Cohen's (1988, 23) formula 2.2.6, $r = d/[(d^2 + 4)^5]$ was used to convert Cohen's d effect size to Pearson's r effect size to clearly understand the range of effect sizes across nine of the studies with the exception of Abrami, Leventhal, and Perry (1982), as well as L'Hommedieu, Menges, and Brinko (1990), due to omega-squared and

Glass's delta being reported, respectively. Converting *d* effect sizes to *r* effect sizes assumes that mean differences are correlational in nature, which results in a variance accounted for effect size (Roberts and Henson 2002) and allows researchers to clearly re-express all effect sizes (Vacha-Haase and Thompson 2004). When examining the relationship between variables, the *r* statistic has been noted as a more general statistic to use (Quintana and Minami 2006), rather than a group difference statistic (e.g. Cohen's *d*).

Comparing the effect sizes across all nine meta-analyses including 193 studies, the range for Pearson *r* values was between .022–.44, suggesting less than a small effect to close to a large effect among the variables of interests and student evaluations of teachers, or alternatively stated, .04–19% of the variance in the variables of interest is accounted for by SETs. To provide a more accurate understanding of the results, the weighted average effect size was calculated based on the nine meta-analyses; the weighted average was determined by multiplying each meta-analysis effect size by its respective number of included studies and then the products were summed and divided by the total number of studies (see Table 2). The effect sizes that were reported for the nine meta-analyses were calculated at either the individual level (i.e. each student rating for instructor was used for analysis) or the group level (i.e. average rating for instructor by students per course section), and in some meta-analyses both individual and group levels were used to calculate the effect size (e.g. Feldman 1993).

Some meta-analyses did not provide the number of participants or sections per study that were originally included in the analysis. Therefore, a weighted mean score could only be determined for the combinations of these studies by using the number of studies in the meta-analyses. The authors did attempt to calculate the weighted mean effect size based on the total sample size used in each meta-analysis by obtaining the original publications that were referenced in the meta-analysis. However, we were unable to locate some individual studies that were used in the meta-analyses (e.g. unpublished dissertations). Including all nine meta-analyses, we calculated a weighted mean effect size of $r = .264$, which is classified as a small-to-medium effect (Cohen 1988). Therefore, 7% of the variance in the variables examined can be explained by SETs. Subsequently, to more clearly understand the effects that each variable had on SETs, we combined the nine studies with similar constructs that were previously identified. Table 2 presents the combined effects, percent of variance accounted for and Cohen's guidelines (see Table 2).

Table 2. Results from combined meta-analyses.

Meta-analysis topic compared with student evaluation of teaching	Number of meta-analyses combined	Total number of studies	Pearson's *r* effect size	Percent of variance explained	Size of effect
Student achievement	6	137	.312	9.73%	Medium
Feedback/Consultation	2	28	.245	6.00%	Small–Medium
Gender of instructor	1	28	.022	.05%	Less than small
overall total	9	193	.260	6.76%	Small–Medium

Note: Size of effects are based on Cohen's guidelines for *r*, which are: .1 = small, .3 = medium and .5 = large.

Discussion

We analysed meta-analyses from three debated areas of research on SETs: their validity, their susceptibility to gender bias and their effective implementation. The authors combined the results from all six available meta-analyses investigating the relationship between SET ratings and student achievement, as measured by final exams' or final-course grades. Our results suggest that there is a medium effect size between student achievement and SET ratings, with measures of achievement explaining approximately 10% of the variance in ratings on SET measures. Previous researchers have examined the validity of SETs through multisection validity studies (e.g. Cohen 1981; d'Apollonia and Abrami 1997; Marsh and Overall 1980) supporting the validity of SET ratings in that students who achieved more in a course experienced the teaching as more effective. For example, d'Apollonia and Abrami (1997, 1203) found that 'more than 45% of the variation in student learning across sections can be explained by student perceptions of instructor effectiveness', suggesting that this percentage can be seen as an estimated validity measure under the proper circumstances. Subsequently, student achievement accounting for 45% of the variance in SETs is rather large based on the current findings of SETs being explained by only 10% of the variance in student achievement. Given this basis for SET validity, our results do seem to support the construct validity of SETs, showing with data synthesised from six meta-analyses that a relationship exists between student achievement and SETs. The combined results from 137 studies on this relationship yielded a medium effect size, suggesting that SETs are related to student achievement and are valid measures of instructor skill and teaching effectiveness. However, it should be noted that achievement was defined differently for various meta-analyses (e.g. final exams' or final-course grades).

The authors were able to combine the results from two meta-analyses investigating the relationship between SETs and different feedback strategies implemented with instructors. Our results suggest that there is a small-to-medium effect size between improvements in teaching (as measured by increases on subsequent SET ratings) and consultative feedback interventions, explaining approximately 6% of the variance. These results echo the findings of the recent methodologically sound meta-analysis by Penny and Coe (2004), whose results suggest a positive relationship between teaching effectiveness and feedback strategies ($d_+ = .69$). Although previous research suggests that any form of feedback (e.g. written feedback based on SET ratings) paired with SET results produces greater effects than SET ratings alone (Cohen 1980), it appears that feedback strategies which actively involve both the instructor and the consultant are more effective in improving teaching (Penny and Coe 2004).

Lastly, we examined the results of a single meta-analysis on the potential for instructor gender to impact SETs. The fact that only one meta-analysis investigating a potential biasing variable on SET ratings was found suggests that researchers focusing on potential biasing variables may further our understanding significantly by using meta-analytic techniques. Although the question of whether the variable 'gender' biases SETs' benefits from one meta-analysis, the authors acknowledge that this study could not synthesise results from many studies and would benefit from additional meta-analyses on the same topic. However, the meta-analysis included in our study yielded a less than small effect size, explaining less than .5% of the variance in SET ratings. Although a controversial issue in student ratings' literature, it appears from the meta-analytic evidence available in the literature that the interaction between instructor gender and student gender does not appear to greatly impact SET ratings.

Teaching implications

Institutions of higher education continually strive to improve the quality of teaching in their academic communities. In addition, administrators at these institutions are constantly trying to increase student achievement. Our results suggest that SET measures can be useful tools in addressing both of these primary concerns. Feedback from student ratings can be useful in improving teaching by providing instructors with constructive, consultative feedback with which to improve the quality of their teaching. As effective teaching has been defined as an instructor's ability to facilitate student achievement (McKeachie, cited in Cohen 1981), using student feedback to improve teaching will likely result in increased student achievement as students benefit from more effective teaching.

One practical way to integrate teaching effectiveness evaluation into an academic class might be to append a SET measure to a course midterm and final examinations. This strategy could provide an effective, practical method to implement student ratings to improve teaching. Some studies examined whether feedback from midterm student ratings was linked with higher student ratings at the semester's end. For example, Overall and Marsh (1979) found that instructors who received midterm student feedback averaged higher on student evaluations at the term's end. Similarly, subsequent researchers (L'Hommedieu, Menges, and Brinko 1990) found an average effect size for improvement in end of term evaluations of .342 for those teachers receiving midterm feedback. With feedback given at a course's midpoint, the instructor would be able to make improvements before the course's end instead of delaying any feedback-driven improvements until the following semester.

Lastly, it appears that interactions between instructor and student gender do not impact SET ratings significantly, according to one meta-analysis. This indicates that instructors should consider neither their gender nor that of their students when receiving and interpreting SET results. The results also suggest that administrators do not need to consider instructors' gender when assigning instructors to various classes. However, the meta-analysis largely included studies from the 1970s, and more current research is needed before making conclusive statements based on gender.

Policy implications

Our results encourage the use of SET measures in evaluating instructor skill and teaching effectiveness. Departmental protocols for evaluating instructors and professors will likely benefit greatly from implementing an SET measure. For those departments currently using a student ratings measure, the authors encourage its continued use and recommend implementing some form of consultative feedback when administratively possible instead of relying on ratings alone. In addition, if a department's SET measure was derived from within the department itself, the department may want to consider ways to validate their measure with other empirically validated scales published in the education or psychology literatures. In selecting an empirically backed assessment, administrators can feel more confident in trusting the results. One SET measure in particular has benefited from ample, sound research and appears to be a reliable and valid, multidimensional measure of teaching effectiveness: the Students' Evaluation of Educational Quality (SEEQ; Marsh 1982). For those departments not currently using an SET scale or for those hoping to improve upon their current measure, the authors recommend implementing an assessment scale, such as the SEEQ, paired with some form of consultative feedback. Furthermore, it should be

noted that most researchers agree with Marsh (2007b) in supporting the idea that SETs ought to comprise only part of the teaching effectiveness evaluation process, especially in making personnel decisions (e.g. Al-Issa and Sulieman 2007; Marsh and Roche 1997; Simione, Cadden, and Mattie 2008; Zabaleta 2007). We are also agreement also with Marsh (2007b), and it is our viewpoint that by using SETs as a single source of data when determining if faculty should be hired, promoted or fired raises great concerns due to the amount of unexplained variance from our study; for example, only approximately 10% of the variance in SETs can be explained by student achievement. Although standard SET measures often provide a simple, inexpensive method for instructor and course evaluation, additional data should be integrated with SETs that are appropriate to the institution. In addition, it is also essential to consider multiple SETs over time, rather than examining only few groups of students/classes.

Feedback strategies

Penny and Coe's (2004) results emphasise that more comprehensive consultation could be incorporated by utilising multiple sources of evaluation in addition to student ratings. From a practical perspective, Penny and Coe's findings indicate that instructors who obtain consultation over student ratings improve their teaching effectiveness, compared with teachers who are provided ratings' results alone. These authors conclude that the most beneficial form of consultation can be characterised as 'collaborative action research aimed at helping teachers to gain a better understanding of the nature of their work and how to improve the quality of their practice in a meaningful way through reflection on their teaching as seen through the eyes of their students' (244). Additional research supports this idea that consultation paired with feedback from student ratings forms an effective way to improve teaching effectiveness (e.g. Hampton and Reiser 2004; Marsh and Roche 1993). We suggest that college and university administrators consider release time for more senior faculty or department chairs to allow those individuals to provide consultation to other faculty members at midterm or semester's end. However, these faculty members should be selected with careful attention and may include a variety of individuals, which may not necessarily be senior faculty. Particularly, academic peers that have demonstrated a consistent and long history of effective teaching would likely be best suited to serve as consultants. By using esteemed academic peers, teachers will likely value the chance to consult with their colleagues who are well known for having the appropriate skills, knowledge and experience (Penny and Coe 2004).

Research implications

Numerous research implications can be extracted from the study. There is a large amount of evidence that supports the validity of SETs. However, we found that multiple variables impact SET ratings differently, and this is an area that should be further explored. One area specifically that warrants further investigation is the biasing variable of gender. In examining the only meta-analysis focusing on gender as the primary biasing variable on SETs, Feldman (1993) concluded that gender had little effect on SETs, but more current research may suggest otherwise (Bachen, McLoughlin, and Garcia 1999; Centra and Gaubatz 2000); it is also worth noting that the majority of studies included in the meta-analysis were conducted in the 1970s. Additionally, Feldman's meta-analysis combined both individual and group analyses,

which could have possibly changed the findings. Researchers should also investigate the effectiveness of SETs on policy decisions, particularly in regard to tenure decisions. We were unable to locate a single meta-analysis that examined how effective SETs were in accurately assisting administrators in their decisions, and future research is needed in this area. At the local level, researchers could investigate the effectiveness of SETs within their university as well as the various strategies that may be employed with SETs (e.g. feedback/consultation). By gathering data within the researchers' university, comparisons with other institutions could be made with published data on SETs.

Limitations

Limitations exist in the present study. Knowing that there are various types of scales that measure student evaluations of teachers, we did not report all of the scales that were used in the meta-analyses because they were not originally reported. This prevents more explicit assumptions about the specific SET scales and how different variables may or may not impact the scales. Therefore, future research should examine the specific and different effects for each SET scale. Another limitation to the study was based on how student achievement was measured differently across studies (e.g. final exams' or final-course grades). Subsequently, it is possible that student achievement may be varied based on the way it was measured; thus, researchers should be consistent when measuring student achievement and clearly operationally define how it is measured, such as student's grade point average at the end of the class.

When provided, confidence intervals were reported, but our study was limited by not reporting confidence intervals for all effect sizes. Future researchers should ensure that confidence intervals are indicated, which allows readers to know if there is a true effect between the variable of interest. Our study was limited by combining only nine of the 11 meta-analyses, as a result of various effect sizes formulas being used. When conducting future meta-analyses, it would be helpful for researchers to provide appropriate descriptive statistics (e.g. *M*; SD) for each study included and the overall sample sizes to allow for further analysis by others. The inclusion criteria added a limitation to the study: due to the file drawer bias, unpublished studies may have impacted our results, but this effect could not be determined. Another limitation to our study was the weighted mean effect size calculations. Unfortunately, the authors were unable to obtain total sample sizes used for each meta-analysis, which prevented us from performing more robust weighted effect size calculations. Where appropriate, in future studies researchers should use Aaron, Kromrey, and Ferron's formula for calculating effect sizes (cited in Vacha-Haase and Thompson 2004) because it tends to be more precise by accounting for small or disparate/unequal sample sizes (Vacha-Haase and Thompson 2004). Surprisingly, the median year for all of the studies included in the meta-analysis was 1976; therefore, a limitation exists based on how dated the majority of the research is in the area.

Lastly, our study suffered from being able to include only one meta-analysis on potential biasing variables. The included paper on the relationship between gender and SETs comprises the beginning of the research necessary to reach an empirically supported decision about whether the interaction between instructor and student gender significantly impacts SETs. However, our study based its conclusions on the available research and noted any weaknesses in approach and conclusions.

Note

1. References marked with an asterisk indicate studies included in the meta-analysis.

Notes on contributors

Stephen L. Wright is an assistant professor in the Department of Counseling Psychology at the University of Northern Colorado. His research interests include attachment, work–family interface, gifted adults and college development.

Michael A. Jenkins-Guarnieri is an advanced doctoral student in the Counseling Psychology program at the University of Northern Colorado. His research interests include the psychosocial development of emerging adults in college, interpersonal competence and online social behaviour.

References

* Abrami, P.C., L. Leventhal, and R.P. Perry. 1982. Educational seduction. *Review of Educational Research* 52: 446–64.

Al-Issa, A., and H. Sulieman. 2007. Student evaluations of teaching: Perceptions and biasing factors. *Quality Assurance in Education* 15: 302–17.

Ambady, N., and R. Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology* 64: 431–41.

Bachen, C.M., M.M. McLoughlin, and S.S. Garcia. 1999. Assessing the role of gender in college students' evaluations of faculty. *Communication Education* 48: 193–210.

Boysen, G.A. (2008). Revenge and student evaluations of teaching. *Teaching of Psychology* 35: 218–22.

Campbell, H.E., K. Gerdes, and S. Steiner. 2005. What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management* 24: 611–20.

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.

Centra, J.A., and N.B. Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *Journal of Higher Education* 71: 17–33.

* 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.

* Cohen, P.A. 1980. Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education* 13: 321–41.

* 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.

* Cohen, P.A. 1982. Validity of student ratings in psychology courses: A research synthesis. *Teaching of Psychology* 9: 78–82.

* Cohen, P.A. 1983. Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education* 54: 448–58.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. New York: Academic Press.

d'Apollonia, S., and P.C. Abrami. 1997. Navigating student ratings of instruction. *American Psychologist* 52: 1198–208.

Dommeyer, C.J., P. Baum, K.S. Chapman, and R.W. Hanna. 2002. Attitudes of business faculty towards two methods of collecting teaching evaluations: Paper vs. online. *Assessment & Evaluation in Higher Education* 27: 455–62.

* Dowell, D.A., and J.A. Neal. 1982. A selective review of the validity of student ratings of teaching. *Journal of Higher Education* 54: 459–63.

* Feldman, K.A. 1993. College students' views of male and female college teachers: Part II – evidence from students' evaluations of their classroom teachers. *Research in Higher Education* 34: 151–91.

Greenwald, A.G., and G.M. Gillmore. 1997. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology* 89: 743–51.

Griffin, B.W. 2004. Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology* 29: 410–25.

Gump, S.E. 2007. Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly* 30: 56–69.

Hampton, S.E., and R.A. Reiser. 2004. Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education* 45: 497–527.

* L'Hommedieu, R., R.J. Menges, and K.T. Brinko. 1990. Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology* 82: 232–41.

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.

Marsh, H.W. 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.

Marsh, H.W. 2007a. Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology* 99: 775–90.

Marsh, H.W. 2007b. Students' evaluations of university teaching: A multidimensional perspective. In *The scholarship of teaching and learning in higher education: An evidence based perspective*, ed. R.P. Perry and J.C. Smart, 319–84. New York: Springer.

Marsh, H.W., and M.J. Dunkin. 1997. Students' evaluations of university teaching: A multidimensional perspective. In *Effective teaching in higher education: Research and practice*, ed. R.P. Perry and J.C. Smart, 241–320. New York: Agathon.

Marsh, H.W., and J.U. Overall. 1980. Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology* 72: 468–75.

Marsh, H.W., and L. Roche. 1993. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal* 30: 217–51.

Marsh, H.W., and L.A. Roche. 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist* 52: 1187–97.

Marsh, H.W., and L.A. Roche. 2000. Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology* 92: 202–28.

* McCallum, L.W. 1984. A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education* 21: 150–8.

Menges, R.J., and K.T. Brinko. 1986. Effects of student evaluation feedback: A meta-analysis of higher education research. Paper presented at the annual meeting of the American Educational Research Association, April, in San Francisco, CA.

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

Overall, J.U., and H.W. Marsh. 1979. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology* 71: 856–65.

* Penny, A.R., and R. Coe. 2004. Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research* 74: 215–53.

Pounder, J.S. 2007. Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education: An International Perspective* 15: 178–91.

Quintana, S.M., and T. Minami. 2006. Research guidelines for meta-analyses of counseling psychology. *Counseling Psychologist* 34: 839–77.

Remedios, R., and D.A. Lieberman. 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.

Richardson, J.T.E. 2005. Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education* 30: 387–415.
Riniolo, T.C., K.C. Johnson, T.R. Sherman, and J.A. Misso. 2006. Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology* 133: 19–35.
Roberts, J.K., and R.K. Henson. 2002. Correction for bias in estimating effect sizes. *Educational and Psychological Measurement* 62: 241–53.
Seldin, P. 1984. Faculty evaluation-surveying policy and practices. *Change* 16, no. 3: 28–33.
Seldin, P. 1993. The use and abuse of student ratings of professors. *Chronicle of Higher Education* 39, no. 46: A40.
Shevlin, M., P. Banyard, M. Davies, and M. Griffiths. 2000. The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education* 25: 397–405.
Simione, K., D. Cadden, and A. Mattie. 2008. Standard of measurement for student evaluation instruments. *Journal of College Teaching & Learning* 5: 45–58.
Vacha-Haase, T., and B. Thompson. 2004. How to estimate and interpret various effect sizes. *Journal of Counseling Psychology* 51: 473–81.
Zabaleta, F. 2007. The use and misuse of student evaluations of teaching. *Teaching in Higher Education* 12: 55–76.