Student evaluations of teaching often play an important role in the promotion and tenure of faculty. Many of the instruments in use within the university setting are self-developed and have not undergone psychometric testing. More specifically, nursing education has many unique challenges, yet the effectiveness of student evaluations of teaching in nursing is understudied. The purposes of this integrative review were to review the current research literature on the design and psychometric testing of teaching evaluation scales, to synthesize the results, and to propose implications specific to nursing programs that are evidence based and will provide guidance on faculty teaching evaluations. A systematic literature review was conducted using Cooper, H. (1998). Synthesizing research: A guide for literature review (3rd ed.). Thousand Oaks, CA: Sage Publications methodology. Data analysis led to the identification of 6 themes: faculty characteristics and perceptions, student characteristics and perceptions, course characteristics, on-line versus paper student evaluations, reliability and validity, and development of new instruments. This literature review provides the context in which to interpret student evaluations of teaching and discusses implications for practice. (Index words: Faculty rating; Student evaluations; Teaching; Course evaluation; Nursing) J Prof Nurs 29:e10–e24, 2013. © 2013 Elsevier Inc. All rights reserved.
“lacked adequate reliability and/or validity data, were too lengthy to encourage voluntary participation, or failed to provide sufficient breadth or depth to meet the college’s instrumentation needs” (Emerson & Records, 2007, p. 3).

Although the effectiveness of SET instruments impacts the careers of nursing educators in significant ways and the research literature on the topic is immense, no summaries of the recent research findings have been published in the nursing literature. The purposes of this integrative review were to review the current research literature on the design and psychometric testing of teaching evaluation scales, to synthesize the results, and to propose implications specific to nursing programs that are evidence based and will provide guidance on faculty teaching evaluations.

**Background**

The most frequent sources of data to document faculty teaching effectiveness include peer review, self-evaluation, administrative evaluation, and student evaluation. Faculty tend to agree that the student evaluations weigh deeply in administrative decision making (Barth, 2008; Beran & Rokosh, 2009b). Unfortunately, many of the instruments used in faculty evaluation systems are “homemade” and, therefore, suspect in their validity and reliability. Moreover, “higher education has yet to establish a universally accepted definition of the characteristics and skills necessary for teaching excellence” (Arreola, 2007, p. 98).

A number of variables influence student evaluations of instruction. Student perceptions are influenced by a variety of fixed and dynamic professor/instructor-held traits, including personality, age, gender, and teaching style (Sprinkle, 2008). Other educators question how the timing of the evaluation may influence the outcomes of student feedback (McNulty et al., 2010). Another theme found in a literature review of the topic included the “leniency hypothesis” that questions whether instructors with more lenient grading criteria may receive more favorable ratings and therefore gain better evaluations (Gump, 2007; Heckert, Latier, Ringwald-Burton, & Drazen, 2006).

Clearly, questions loom concerning the practical issues in obtaining valid student feedback. These include issues of use of formal instruments, the kinds of feedback collected, the timing and data collection methods of the feedback, response rates, and the usefulness of a single questionnaire for different types of students, including graduate students and distance-learning students (Richardson, 2005).

Nursing education programs bring unique challenges to faculty evaluation systems because a variety of types of faculty and instructional settings are employed. These include simulation laboratories, health assessment laboratories, classroom, on-line or blended teaching, and diverse settings for practicum experiences. Yet, few studies in the nursing literature have looked specifically at student evaluations of nursing instruction or proposed any evidence-based implications for improving student evaluation methods in nursing education. Clearly, a new wave of research is needed that critically and holistically appraises past studies with respect to their methodologies, conclusions, and implications (Gump, 2007). This integrative review aims to fill in the gap in the nursing literature.

The terms student evaluations and faculty are not clearly defined within the literature. The dictionary definition of student is “one who attends school” with evaluation defined as “to determine the value of” (Merriam-Webster Online, 2011). Thus, for the purposes of this literature review, “student evaluations” is defined by the authors as “a determination of the value of teaching in a college course by one who is enrolled in the course.” “Faculty” is defined as “the teaching and administrative staff and those members of the administration having academic rank in an educational institution.” (Merriam-Webster Online, 2011). For this literature review, the authors define “faculty” as teachers having academic rank in an institution of higher education.

**Methodology**

A systematic literature review was conducted utilizing the following databases: Cumulative Index of Nursing and Allied Health Literature, Educational Resources Information Center, PubMed, and PsycINFO. The search included the following terms that were searched both in singular and in a Boolean search with the term AND for several combinations in “abstracts”: student evaluations, teaching, evaluations, teacher performance, faculty development, course evaluation, job performance, faculty, faculty rating, and nursing. The complete search explored literature from 2001 to 2011 with the delineation of results limited to “research articles only.” Initially, the search was limited to publication dates within the past 5 years. This was expanded back to 2001 when initial results were limited.

The search was guided by Cooper’s (1998) integrative review methodology as a framework to enhance methodological rigor. Cooper (1998) developed five stages for research synthesis: (a) problem formulation, (b) literature search, (c) data evaluation of the quality of the research studies, (d) analysis and interpretation, and (e) presentation of the results. The variables of interest for this literature review were student evaluations of teacher performance and courses at the undergraduate or graduate level.

Criteria for inclusion in this integrative review included (a) publication dates between 2001 and 2011, (b) published in English, (c) qualitative and quantitative studies that were published in peer-reviewed journals, and (d) the study reports the method of data collection and analysis even if descriptions are incomplete. Letters to the editor, publications not representing original research, and unpublished manuscripts (abstracts and dissertations) were excluded from this integrative review. In addition to a systematic review of the stated databases, an ancestry search of selected articles from the initial
search was performed to identify additional studies that addressed the variable of interest. Data were extracted from primary sources only.

The search terms provided an extensive listing of potential studies. All article titles were reviewed for possible relevance to the concept of student evaluation of teacher performance and/or courses at the undergraduate or graduate level. Abstracts of articles with relevant titles were then reviewed. If the abstracts met inclusion criteria, the full studies were evaluated for inclusion in the literature review analysis.

Coding sheets were developed to outline each article. A minimum of two authors reviewed each study and verified the coded content. After analyzing for emerging themes, the studies were then alphabetically categorized according to author's last name.

Fifty-one abstracts were identified and examined. After a thorough evaluation of each study, 20 articles were excluded because their primary focus was not specific to our question (e.g., faculty evaluated by administration or focused only on adjunct or preceptor faculty). Thirty-one articles were retained for the final analysis and interpretation. Only 7 of these articles were related specifically to nursing (see Table 1).

Of the 31 articles included in the review, the majority (n = 26) used quantitative methods only. Two of the included studies used only qualitative methods and 3 used mixed methods. Articles were categorized by an evidence hierarchy schema by Melnyk and Fineout-Overholt (2011), which demonstrates a seven-level hierarchy that has at its pinnacle systematic reviews of randomized clinical trials. All of the studies included in this review had a level of evidence of six. According to the Melnyk and Fineout-Overholt definition, Level 6 includes single descriptive or qualitative studies (Melnyk & Fineout-Overholt, 2011). Most of the studies were conducted in either the United States or Canada. A few studies did not describe their location. The four other studies were conducted in Beirut (El Hassan, 2009), Taiwan (Tang, Chou, & Chiang, 2005), India (Alok, 2011), and Australia (Salamonson, Halcomb, Andrew, Peters, & Jackson, 2010).

The study population for most (25 studies) of the articles was either students or student evaluations, although 11 studies examined faculty. Two studies included administrators as well. The majority of the studies that used students for their population examined undergraduates only.

Results/Themes

Faculty Characteristics and Perceptions

Several faculty characteristics are related to positive SET. The ability to clearly teach material and to be organized were the most common faculty characteristics noted to be associated with good evaluations (Donnon, Delver, & Beran, 2010; Sitzman, 2010; Wolf, Bender, Beitz, Wieland, & Vito, 2004). The need for clarity was important for courses held on-line (Sitzman, 2010). The amount of teaching experience or expertise of the faculty member was generally related positively with higher rankings on evaluation instruments (Nowell, Gale, & Handley, 2010; Sitzman, 2010). Part-time ("sessional") staff received higher student ratings from undergraduates than tenured faculty, but this trend is reversed for upper-class students (Salamonson et al., 2010). Faculty seen by students as friendly, enthusiastic, fair, and helpful were rated as more effective instructors than faculty who were not (Barth, 2008; Hills, Naegle, & Bartkus, 2009; Tang et al., 2005). One study found that White faculty scored significantly higher than Black faculty on multidimensional items and on "overall value of course" and "overall teaching ability" (Smith & Hawkins, 2011). Conversely, lack of organization, lack of clarity in teaching, poor time management, the rigor of the course, and inaccessibility were common faculty weaknesses linked to lower SET (Barth, 2008; Sitzman, 2010; Wolf et al., 2004).

There seems to be a divide among faculty in studies that examined faculty perceptions of the student evaluation process. Some studies noted positive faculty views of the student ratings process. Most of the faculty believe that SET are meaningful and report that they made improvements based upon them (Balam & Shannon, 2010; Beran & Rokosh, 2009b; Beran & Violato, 2005; Beran, Violato, Kline, & Frideres, 2005; El Hassan, 2009). The more positively faculty view ratings, the more likely faculty are to report that ratings are useful (Beran & Rokosh, 2009a; Beran & Rokosh, 2009b).

Conversely, other studies noted that faculty are skeptical of the validity of the process and the efforts that students put into their evaluations of faculty teaching. Many faculty believe that student ratings are unreliable and not valid (Balam & Shannon, 2010) and that they do not have much of an impact on enhancing instruction (Beran & Rokosh, 2009a; Beran & Rokosh, 2009b). Most of the faculty (60%) did not believe that students took the evaluation process seriously (El Hassan, 2009). Further, faculty believe that students are not experts in pedagogy and, thus, can only assess their own response to a course (Ackerman et al., 2009).

Student Characteristics and Perceptions

Several student characteristics were associated with high student evaluations of instructors. Students who expected a high grade were more likely to provide higher ratings (Barth, 2008; Beran & Rokosh, 2009b; Beran & Violato, 2005; Beran & Violato, 2009; Donnon et al., 2010; Nowell et al., 2010; Serdyukova, Tatum, & Serdyukov, 2010; Sonntag, Bassett, & Snyder, 2009). In addition, students who attended class provided the highest ratings of instructors (Beran & Violato, 2005; Beran & Violato, 2009; Hills et al., 2009). Beran and Violato, (2005) however, report that higher grades only account for a small portion of the variance among ratings. A curvilinear relationship was found between grades and SET, whereby the best and worst students in a course report the most favorably on SET (Serdyukova et al.,...
<table>
<thead>
<tr>
<th>Author/Citation</th>
<th>Sample</th>
<th>Study purpose</th>
<th>Variables</th>
<th>Study methods and instruments</th>
<th>Study outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ackerman et al. (2009)</td>
<td>Business or finance faculty in undergraduate and graduate programs Faculty interviews: n = 8</td>
<td>Examine the views and perceptions of university faculty members about SET and peer observations of teaching as means of measuring teaching effectiveness.</td>
<td>Peer observation reports SET</td>
<td>Qualitative Exploratory inquiry using minimally structured in-depth interviews. Analysis using iterative process of coding, categorizing, and abstracting data to identify themes.</td>
<td>Evaluation of teaching effectiveness should incorporate both SET and peer observations of teaching for optimal evaluation. Faculty experts in evaluating course content; students are experts in assessing their response to a class.</td>
</tr>
<tr>
<td>Alok (2011)</td>
<td>Centurion School of Rural Enterprise Management, India Instructor and student interviews.</td>
<td>Faculty and student development of a student evaluation of teaching instrument</td>
<td>16 key areas along three performance dimensions (course organization, quality of teaching, and assessment and feedback).</td>
<td>Cronbach’s alpha Principal component analysis Varimax rotation factor analysis</td>
<td>Development of SET. 18 items generated on a 6-point Likert with two factors of learning orientation and learner orientation. Qualitative response area.</td>
</tr>
<tr>
<td>Balam and Shannon (2010)</td>
<td>Students: n = 968 Faculty: n = 34</td>
<td>Analyze differences between college students’ and faculty’s perceptions, based on the student ratings myths by Aleamoni (1987, 1999)</td>
<td>Gender, faculty and student attitudes toward myths, student ratings, GPA, tenure status, Student Evaluation of Educational Quality (SEEQ), college/school, professional rank</td>
<td>Results were compared using multivariate analysis of variance tool: SEEQ Regression analysis to compare student characteristics with their perceptions of student rating myths</td>
<td>Faculty found that student ratings are largely unreliable and not valid. Neither students nor faculty subscribed to the myths</td>
</tr>
<tr>
<td>Barth (2008)</td>
<td>Average class SET scores from 167 classes over 3 years from Georgia Southern University (GSU). Students: n &gt; 4,000</td>
<td>Determine which aspects of the GSU SET have the most impact on overall instructor rating.</td>
<td>Quality of instruction, course rigor, level of instruction interest, grades, instructor helpfulness, overall instructor rating</td>
<td>Factor analysis with two steps: (a) develop a set of measures on the traits that are latent in the instrument(b) use multiple regression analysis to evaluate how those traits affect Questions 18 (overall instructor rating)</td>
<td>Original variables grouped into five factors (explanatory variables): quality of instruction, course rigor, level of interest, grades, and instructor helpfulness were identified. Primary driver of overall instructor rating is quality of instruction. Negative impact of rigor can be overcome with instructor helpfulness.</td>
</tr>
<tr>
<td>Beitz and Wieland (2005)</td>
<td>Northeastern U.S., university students n = 198</td>
<td>Compare full- and part-time basic bachelor of science in nursing (BSN), licensed professional nurse—BSN, and registered nurse—BSN students’ ratings of effective clinical teaching behaviors (ECTB)</td>
<td>Full time versus part time, type of program</td>
<td>Nursing Clinical Teaching Effectiveness Inventory—a 48-item Likert scaleECTB inventory—a 43-item, 5-point Likert scale</td>
<td>No statistically significant difference in ratings of teaching effectiveness based upon type of program. Part-time students rated their clinical faculty at a statistically significantly higher level than full-time students.</td>
</tr>
<tr>
<td>Beran, T.N. &amp; Rokosh, J.L. (2009b)</td>
<td>Canadian university faculty Faculty: n = 357</td>
<td>Examine attitudes about student ratings of instruction and faculty attitudes</td>
<td>Faculty demographics and use of ratings Faculty attitudes re: utility for instruction versus administrative evaluation</td>
<td>Analysis of Variance and Pearson Product Moment correlations to determine if instructor attitudes differed according to variables in USRI</td>
<td>Faculty believe that ratings are helpful to rate institutional integrity and is useful to administrators in decision making. Faculty do not believe that ratings have that much of an impact on enhancing instruction</td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Author/Citation</th>
<th>Sample</th>
<th>Study purpose</th>
<th>Variables</th>
<th>Study methods and instruments</th>
<th>Study outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beran, T.N., &amp; Rokosh, J.L. (2009a)</td>
<td>Canadian university faculty Faculty: n = 357</td>
<td>Examine the validity of student ratings to determine the extent of improving teaching effectiveness.</td>
<td>Usefulness of instrument (USRI) ratings Demographic data</td>
<td>Survey No statistical analyses discussed.</td>
<td>67% expressed a negative view 70% expressed concern with the USRI instrument itself (inadequate and provides little or no assistance in instructional improvement, lack of qualitative feedback) USRI was not appropriate means of evaluating Web-based, team-taught courses. Unfair to compare one faculty with another.</td>
</tr>
<tr>
<td>Beran and Violato (2005)</td>
<td>Canadian University students Students: n = 371,131</td>
<td>Examine the importance of student and course characteristics in relation to student ratings.</td>
<td>Course status (workload, required/elective, departmental) Course description (type and length) Student engagement (attendance, expected grade)</td>
<td>Tool: USRIANOVA, regression analysis</td>
<td>Most significant predictor of course rating was student's expected grade. Class attendance was related to student ratings. Students give high ratings to instructors that they consider effective. Course characteristics do not explain how students rate their instructors.</td>
</tr>
<tr>
<td>Beran and Violato (2009)</td>
<td>Canadian University students Students: n = 371,131</td>
<td>Identify interrelationships of course characteristics, student engagement, and student ratings of instruction</td>
<td>Course status (workload, required/elective, departmental) Course description (type and length) Student engagement (attendance, expected grade)</td>
<td>Quantitative Two-step analytic procedure Latent variable path analysis model Tool: USRI</td>
<td>Course characteristics not directly related to student ratings but indirectly through student engagement. Student engagement related to student ratings; students expecting a higher grade, and attending class regularly rated faculty higher. Compared to lectures, applied courses (laboratories, practica) did not have higher ratings.</td>
</tr>
<tr>
<td>Beran et al. (2007)</td>
<td>Canadian University administrators Administrators: n = 52</td>
<td>Survey of administrators about the types of ratings they use, how useful they are, and their purpose.</td>
<td>Instructor characteristics (promotion and tenure) Course characteristics (course timetabling) Student rating usefulness</td>
<td>Likert survey analyzing the USRI; principal component analysis with varimax rotation performed Pearson Product Moment to examine purpose and administrative utility Content analysis for open-ended question</td>
<td>Administrators: interested in knowing about instructor characteristics/teaching style; use ratings for instructor/department evaluation and scheduling courses; ratings are insufficient to be used alone in evaluation of faculty but often are</td>
</tr>
<tr>
<td>Beran et al. (2009)</td>
<td>Canadian University students Students: n = 1229</td>
<td>Measure students' perspectives on the use of student rating information available to university students and to quantify the utility of student ratings of</td>
<td>Student characteristics Instructor characteristics Course characteristics Instructor's relative ranking</td>
<td>Twelve items rated on a 4-point scale Factor analysis with subsequent oblique rotation MANOVA—used to determine if instructor characteristics, course materials, and rating</td>
<td>Students use ratings primarily to select courses. The factors of student ratings' usefulness (instructor characteristics, course materials, and instructor's relative ranking) vary according to the characteristics of the users. Young and</td>
</tr>
</tbody>
</table>
Beran et al. (2005). Students, faculty, and administrators at a Canadian University: \( n = 1229 \)  

Investigate the utility or consequential validity of student ratings of instructors varied with student characteristics. Wilk’s Lambda —group differences for student characteristics  

Use of ratings for future student course selection Faculties use of student ratings to improve teaching Administrators use of ratings about decisions of merit, raise, or promotion  

Survey questions were administered to faculty, students, alumni, and administrators at the third year postimplementation of a new student rating system tool (Universal Student Rating of Instruction Instrument)  

Students use the ratings to assist them in determining which class or instructor to take in the future; faculty use the ratings to sometimes change their teaching; administrators use to assist in decisions on promotion, tenure, and salary adjustments.

Brown et al. (2009) NY college Students: \( n = 110 \) Faculty: \( n = 312 \)  

Examine how students perceive and use RMP ratings. Determine if RMP ratings are valid predictors of instructor’s ability.  

Teachers’ ability to communicate clearly Teachers’ availability to students outside of class Rating of the difficulty of tests or examinations in said course.  

Exploratory survey of students compared RateMyProfessors.com ratings with course evaluations from the same instructors on the three variables selected.  

Students use RMP when making academic decisions and believe RMP ratings are honest. RMP ratings are significant predictors of instructors performance on course evaluations (communication, availability, difficulty of examinations).

Catano and Harvey (2011) Senior-level university classes in the arts, science, and business “subject matter experts:” Undergraduate students = 65 Graduate students = 8  

Develop a SET to establish effective job performance measurement systems.  

Nine categories in the new scale: communication, availability, creativity, individual consideration, social awareness, feedback, professionalism, conscientiousness, and problem solving.  

Used undergrads to create critical incidents. Used graduate students to rate the incidents (7-point scale) Means and standard deviations, factor analysis, Cronbach’s alpha, correlations  

One factor created in factor analysis with all nine categories falling within it—68% of the variance accounted for. The new tool (ETCS) is reliable and valid measure of teaching effectiveness. The ETCS is short and concentrates on factors within the instructor’s control.

Clayson and Haley (2011) Marketing business students: \( n = 236 \)  

Conduct a literature review of SET. To conduct a questionnaire exploring the proportion of students who admit to giving purposeful misinformation and know of other students who provide false information.  

Student opinions of falsification of SET information  

Confidential 21-question survey/questionnaire, completed in classes at two American universities. Validity check between universities.  

Most of the (56%) students knew of other students who deliberately falsified SET. An estimated 30% of SET contain answers that students know to be untrue.

Donnon et al. (2010) Canadian University Student rating forms; \( n = 1738 \)  

Course characteristics (workload; required or elective) Expected course gradeStudent rating of quality of course instruction on Likert scale  

Mandatory courses had a significantly lower mean rating of overall quality of instruction. Higher course workload had lower mean ratings for overall quality of instruction. Students who expected to get an “A” provided significantly higher overall quality ratings. The most significant predictor was presenting...
<table>
<thead>
<tr>
<th>Author/Citation</th>
<th>Sample</th>
<th>Study purpose</th>
<th>Variables</th>
<th>Study methods and instruments</th>
<th>Study outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>El Hassan (2009)</td>
<td>American University of Beirut Students: n = 605 Faculty: n = 145</td>
<td>Identify consequences of using evaluations to understand the process students use to respond to evaluations and how faculty use the evaluation.</td>
<td>Student perceptions Faculty perceptions</td>
<td>t tests—mean differences of uncorrelated dataTwo surveys (one for students and one for faculty)—12 items each evaluating the Instructor Course Evaluation form; 5-point rating scale plus open-ended comments</td>
<td>Course material in a well-organized manner to achieve high quality ratings. Students and faculty believe in the effectiveness and usefulness of the evaluation process. Students believe ratings affect what is addressed in class. Faculty see ratings as positive, except for personnel decisions (salary, promotion)</td>
</tr>
<tr>
<td>Emerson and Records (2007)</td>
<td>Pilot: tested evaluation tool: 13 courses Second phase: N = 148 from 10 didactic courses N = 149 from 13 courses</td>
<td>Student course evaluation tool development—driven by faculty survey, pilot tested, full psychometric testing for validity/reliability in undergrad didactic and clinical courses, grad didactic courses</td>
<td>Student evaluation tool development Five domains:-Communication, -Expertise, -Professionalism, -Advocacy, -Pedagogy</td>
<td>Construct validity via exploratory factor analysis Cronbach's alpha for reliability</td>
<td>Results were robust for courses evaluated Majority of students believed that instructors were teaching effectivelyIndicates need for multiple data sources for evaluating teaching effectiveness</td>
</tr>
<tr>
<td>Frick et al. (2010)</td>
<td>A large midwestern university 464 students in 12 courses &quot;Nearly all were&quot; undergraduate</td>
<td>Examine the relationships between study variables. Compare student responses with Teaching and Learning Quality tool, which incorporates use of first principles of instruction (FPI) and academic learning time (ALT). Examine whether success in course activities related to mastery of course objectives.</td>
<td>ALT, course satisfaction, learning progress, instructor and course quality, student mastery of course objectives.</td>
<td>TALQ toolSpearman correlations Analysis of patterns in time Hierarchichal loglinear model Factor analysis</td>
<td>If students agreed that their instructors used FPI and agreed that they experienced ALT, then students were about five times more likely to achieve high levels of mastery of course objectives and 26 times less likely to achieve low levels of mastery.</td>
</tr>
<tr>
<td>Guder and Malliaris (2010)</td>
<td>Loyola University, Chicago, School of Business 17,161 student online and paper evaluations completed over 4 years (eight semesters).</td>
<td>Compare the response rate, instructor and course rating, and student comments provided by students on online versus paper SET</td>
<td>Response rateInstructor and course rating numbers Number and length of comments provided</td>
<td>Percentage change in response rates Correlations between response rates and the three variables (class size, instructor rating, and class rating)</td>
<td>Twenty-five percent drop in responses to the SET when changed from paper to online. No significant drop in instructor or course ratings was noted. More students provided comments on the on-line form, and students provided lengthier comments on the on-line form.</td>
</tr>
<tr>
<td>Reference</td>
<td>Description</td>
<td>Methodology</td>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hill &amp; Epps (2010)</td>
<td>Analyze the impact of classroom environment factors on individual student satisfaction measures and SET in the university environment.</td>
<td>Classroom (old versus upgraded)Time of day class taughtSize of class Instructor Expected course grade Demographic information</td>
<td>Chi-square tests regression analysis Differences affected the students’ perceptions of the instructors’ organization, their own enjoyment of the class, their perceived level of learning, and their general sense of satisfaction. Students in the upgraded classroom perceived the instructor to be more organized, and they felt more strongly that they learned something new each class in the upgraded classroom.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hills et al. (2009)</td>
<td>Examine students’ degree of salience with regard to the typical course evaluation.</td>
<td>Gender, class year, GPA, college affiliation (business or nonbusiness majors vs. undeclared).</td>
<td>Varimax-rotated factor analysis Descriptive statistics Cronbach’s alpha reliability Factor analysis = course traits, instructor traits, and class participation. Items rated as highly important: fairness of grading, relevant materials, and examinations, using class time to help students learn the course content. Items rated as less important: ability to participate actively in class, appropriateness of course readings.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nowell et al. (2010)</td>
<td>Examine disparity between on-line and in-class SET ratings.</td>
<td>Instructor quality (organization, willingness to respond to students, availability, respect for students, overall contribution of instructor), expected grade, class size, class level# class/week, Full-time versus part-time faculty Faculty years of teaching Perceptions of Teaching and Course Satisfaction (PTCS) scores Sessional versus tenure/FT faculty Students assessments (grading)</td>
<td>Multiple regression analysis Many student and class characteristics not under the control of the instructor—if controlled, the average SET ratings on-line are significantly lower than in-class ratings. Grades and student ratings are positively correlated. Class size increase, SETs significantly decline. Expected grade and years teaching directly related to higher SET rankings.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salamonson et al. (2010)</td>
<td>Examine differences in (a) student ratings of their satisfaction with the teaching of sessional and tenured staff and (b) assessment scores awarded by sessional and tenured staff.</td>
<td>PTCS scale Kaiser–Meyer–Olkin Chi-square Cronbach’s alpha reliability</td>
<td>Report tool is reliable and valid. On-line classes receive lower evaluations than onsite classes. On-line classes receive lower GPAs than onsite classes. Two factors: the factors suggest that the instrument measures, teaching ability, and some combination of learning and course</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serdyukova et al. (2010)</td>
<td>Examine the reliability and validity of a revised SET tool and to do a comparative analysis across types of classes (on-line, onsite, etc.), student level (undergraduate Course, month taught, school/college, department, type of class, campus, level (undergraduate vs. graduate), enrollment, response rate, faculty rank, NU’s SET assessment form—a revision, 30 items Cronbach’s alpha Comparative analysis Factor analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author/Citation year</td>
<td>Sample</td>
<td>Study purpose</td>
<td>Variables</td>
<td>Study methods and instruments</td>
<td>Study outcomes</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------</td>
<td>---------------</td>
<td>-----------</td>
<td>-------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Sitzman (2010)</td>
<td>Students from five universities in on-line courses Students: n = 122</td>
<td>vs. graduate), different schools/colleges, and faculty ranks.</td>
<td>class GPA, and a rating for each of the 30 items.</td>
<td>Mixed methods Survey that prioritized previously identified caring instructor behaviors. One open-ended question about caring behaviors—analyzed themes.</td>
<td>Clarity/Expertise associated with teachers' ability to convey academic content and expectations clearly. Students preferred instructor who could manage on-line process and content/material effectively. Trust considered a thread.</td>
</tr>
<tr>
<td>Smith and Hawkins (2011)</td>
<td>Southeastern U.S. university, college of education13,702 student (grad and undergrad) evaluations 190 tenure-track faculty</td>
<td>Assess the relative importance of caring behaviors of on-line instructors identified by participants; tool development</td>
<td>Student perceptions “Caring behaviors”</td>
<td>36-item course evaluation form “descriptive and inferential statistics” (mean, SD), t test, Cohen’s d</td>
<td>Faculty had higher average mean scores on teaching for most of the items (4.0 or higher on 5-point scale). Of the three groups of faculty (White, Black, and “other”), faculty identified as other received the highest, whereas Black faculty received the lowest mean scores on the 26-multidimensional items and the two global items; Black faculty received higher mean scores on multidimensional items but lower mean scores on the two global items.</td>
</tr>
<tr>
<td>Sonntag et al. (2009)</td>
<td>University faculty Faculty: n = 126</td>
<td>Describe the undergraduate student ratings of teaching effectivenessUse critical race theory to compare the teaching effectiveness for tenure-track faculty based on race</td>
<td>Raceltems on the scale (global and multidimensional)—for example, organization, preparation, or interest in subject matter, overall value of course, and overall teaching ability</td>
<td>Average ratings of Lander University professors from RateMyProfessor.com, IDEA student ratings of teaching, and average assigned course grades. Descriptive statistics, correlations</td>
<td>IDEA and RateMyProfessor.com ratings of teaching quality were significantly correlated with grades. This finding replicates previous research that showed that both traditional and public Web-based teaching evaluations reflected bias to prefer grading leniency. Results provide preliminary support for the validity of RateMyProfessor.com evaluations.</td>
</tr>
<tr>
<td>Tang et al. (2005)</td>
<td>Two Taiwan schools of nursing Students: n = 214</td>
<td>Determine which of the qualities of nursing faculty is the main contributor to effectiveness, as perceived by nursing students</td>
<td>Teacher effectiveness (professional competence interpersonal relationship personality characteristics teaching ability)</td>
<td>Two pilot studies used Cronbach’s coefficient alpha which resulted in a 40-item questionnaire, a 5-point Likert-type scale, analyzed using t tests. Pearson correlation was used to analyze the consistency between the two schools.</td>
<td>Effective teachers rated high in all categories. Ineffective teachers rated low in all categories. Largest differences between effective and ineffective teachers related to interpersonal relationship and personality characteristic categories. Teachers’ attitudes toward the students are considered more influential in student perception than professional ability.</td>
</tr>
</tbody>
</table>
2010). Hills et al. (2009) found that students place the most emphasis on the in-class experience but rank their own active participation in the course low in importance. Age, gender and experience may also have an effect on student ratings of faculty. Undergraduates are more critical of faculty performance than graduates (Wolf et al., 2004) and junior students perceived their instructors more critically than senior students (Wade & Kasper, 2006). Freshmen and sophomores were more concerned with grading, appropriateness of workload, and relevance of course materials, which are suitable to success in typical lecture-based courses, whereas juniors and seniors were more likely to place importance on in-class discussion (Hills et al., 2009). According to Hills, Naegle, & Bartkus (2009), “students value transparency in their courses”; they “valued knowing why they should come to class and why the material they were learning was important” (p. 302).

Beitz and Wieland (2005) found that, although nursing program type did not seem to affect ratings, part-time students rated faculty higher than full-time students. Female students tend to score differently on SET providing a greater range of scores (Hills et al., 2009) and higher scores than their male counterparts (Catano & Harvey, 2011). On-line classes received lower evaluations than traditional courses (Serdylkova et al., 2010).

Students believe that they are able to make accurate assessments of faculty teaching (Balam & Shannon, 2010). However, many students do not perceive the use and function of course evaluations in the same manner as faculty. One fourth (26%) of students agreed with the statement that student ratings of faculty are a meaningless activity and that they performed the activity (SET) only because it is required (El Hassan, 2009). Almost one third (31%) of students believed that a perfect 5 on the rating scale was unattainable (El Hassan, 2009). Further, El Hassan (2009) found that more than half (53%) of students put a 3 on a 5-point scale when they are either undecided or uninterested in the student ratings of faculty. Almost one third (31%) of students reported giving a faculty member a higher or lower evaluation than the instructor deserved, and 19% reported writing something untrue while completing the SET (Clayson & Haley, 2011).

Student behaviors can have a significant negative impact on faculty evaluations. A further consideration is the finding by Clayson and Haley (2011) that 11% of participants did not follow instructions correctly on at least one part of their questionnaire even though instructions were repeated three times. This suggests that errors may also be occurring when students are completing SET, leading to inaccurate information.

Course Characteristics

The characteristics of courses themselves, separate from teaching ability, may have an effect on faculty course evaluations. Mandatory courses had lower student ratings than courses where there was a choice involved (e.g., electives; Donnon et al., 2010). The larger the class size,
the lower the student evaluation scores (Guder & Malliaris, 2010; Nowell et al., 2010). Courses with higher workloads also had lower student rating scores (Donnon et al., 2010). Instructors can modulate this effect, however, if they offer additional assistance to their students (Barth, 2008). Classroom environment had an effect on SET. Students in upgraded classrooms felt that faculty were more organized and were more likely to believe that they were learning something new each day, compared with students in rooms with poor seating, poor lighting, noise issues, and general discomfort (Hill & Epps, 2010). This effect held true even though the course professors in this study taught the classes using the same syllabi, tests, home assignments, reading assignments, and lecture notes (Hill & Epps, 2010). Although one study did not find that course characteristics were directly related to student ratings of faculty but rather that the course characteristics are mediated through student engagement (Beran & Violato, 2009). Many of these course characteristics lay outside of faculty control (Nowell et al., 2010) yet can have a significant effect on teaching evaluation scores.

Online Versus Paper SET
Researchers have also examined the role of on-line versus the traditional paper student evaluation forms. On-line independent venues for students to evaluate their instructors, such as RateMyProfessor.com, allow students to do so publicly. Students believe that RateMyprofessor ratings are honest (Brown, Baillie, & Fraser, 2009). Further, some studies have found that the Ratemyprofessor ratings are valid when compared with traditional course evaluations (Brown et al., 2009; Sonntag et al., 2009). These studies, however, only examined the perception of how easy a course was and the overall quality of a course in comparing RateMyProfessor.com to university-administered evaluations (Brown et al., 2009; Sonntag et al., 2009).

The switch from paper to on-line SET has had a significant negative effect on response rates, which dropped by 25% after on-line SET began (Guder & Malliaris, 2010). Courses with smaller class size had better response rates than courses with a larger class size (Guder & Malliaris, 2010). Findings are conflicted regarding the effect of on-line SET on instructor and course ratings. One study found that the completion of required student evaluations of faculty on-line as opposed to paper-and-pencil evaluations has a negative influence on rating scores (Nowell et al., 2010). Guder and Malliaris (2010), however, found no significant change in SET results after implementation of on-line course evaluations except that students made 19% more comments and 149% longer comments on-line.

Reliability and Validity of SET
Many SET instruments being used for SET lack reliability and validity data (Emerson & Records, 2007). Some of these scales have perceived validity, that is, faculty and students believe them to be valid. Beran and Rokosh (2009b) reported that faculty believe that student evaluations are useful to improve the quality of their teaching. Faculty being evaluated are observed on multiple occasions over the semester giving SET-perceived validity, argue Ackerman et al. (2009). The Universal Student Ratings of Instruction (USRI), which was used in most of the Beran et al. studies, had high internal consistency (reliability; Donnon et al., 2010). Their examination of USRI's construct validity revealed a one-factor explanatory model in which the item “The course material was presented in a well-organized manner” explained 58% of the variance of the global rating of overall quality of instruction (Donnon et al., 2010). A course evaluation instrument used by Serdyukova et al. (2010) was also found to be reliable and valid.

Two studies examined RateMyProfessor.com for reliability and validity comparing them to university rating scales. On-line evaluations were strongly correlated with “able to communicate clearly” on the university ratings, and “helpfulness” was moderately correlated with “availability outside of class” on the university ratings. However, how easy the course was perceived was only slightly correlated with “rate the difficulty of exams in the class” (Brown et al., 2009). The second study found that easiness on the on-line scale was positively correlated with actual grade point average (GPA) assigned for the course and that overall quality on the on-line scale was positively correlated with an excellent instructor (Sonntag et al., 2009). These studies conclude that Internet-based evaluations through RateMyProfessor.com are reliable and valid.

Overall, however, there has not been sufficient examination of reliability and validity of instruments being used to evaluate faculty, many of which are created by an individual university or by a university department. Administrators tend to focus on the “overall rating” SET item as the basis for faculty evaluation (Barth, 2008) even though they admit concerns about the validity of SET (Beran, Violato, & Kline, 2007). These concerns include effects on grade inflation and that evaluations may place too much burden on faculty, rather than students, for student learning (Beran et al., 2007).

Development of New Instruments
Because of persistent concerns regarding the reliability and validity of SET, some have attempted to create newer and better rating scales. Tang et al. (2005) modified an existing instrument to develop a scale that differentiates effective and ineffective teachers. Four studies documented the creation of new SET instruments: the Classroom Instruction Evaluation Scale (Emerson & Records, 2007), the Evaluation of Teaching Competency Scale (ETCS; Catano & Harvey, 2011), and a Teaching and Learning Quality (TALQ) scale (Frick, Chadha, Watson, & Zlatkovska, 2010). All four of these instruments were reported to be reliable and valid. In contrast, Beran, Violato, Kline, and Frideres (2009) utilized an instrument that assessed the usefulness of student evaluations of faculty to students. Only one study examined nursing
students. Wade and Kasper (2006) developed a reliable and valid measure of nursing student’s perceptions of instructor caring.

Each of these scales identified factors, domains, or categories pertinent to faculty and course elements. For example, Emerson and Records’s (2007) five domains include communication, expertise, professionalism, advocacy, and pedagogy, whereas Tang et al.’s (2005) instrument development process revealed four categories: professional competence, interpersonal relationships, personality characteristics, and teaching ability. Several factors were important to teaching effectiveness (communication, availability, creativity, individual consideration, social awareness, feedback, professionalism, conscientiousness, and problem solving) according to students (Catano & Harvey, 2011).

Three of these new instruments used an even-numbered Likert scale, which prevents students from giving a neutral answer (Alok, 2011; Emerson & Records, 2007; Wade & Kasper, 2006). Generally, these instruments were created through interviews with faculty and others and by relevant literature reviews. According to Alok (2011), participation by both faculty and students is necessary for SET development and implementation to be successful.

Discussion

There is a relative scarcity of research about SET when one considers the importance that is placed on SET by students, faculty, and administrators. The highly variable nature of teaching itself makes designing quality case-controlled and randomized studies difficult. SET are fraught with faculty questions of reliability and validity, and although faculty know that there are characteristics about courses, faculty, and students that influence SET, many faculty do not know exactly what these characteristics are. Faculty are frequently autonomous in the teaching setting so that it becomes possible for the students to be the sole witnesses to an instructor’s teaching style and content. SET, with all of their flaws, have become a major component of decisions about tenure and promotion. The emphasis on SET to the exclusion of other evaluative measures can lead to faculty frustration with the system because it currently stands in many university settings.

In this section, the content areas of this literature review (faculty, student, and course characteristics, online format, and reliability/validity) will be discussed to elucidate results and clarify the varying aspects of SET. Several new instruments will be discussed, and implications for practice in nursing education will be given.

Faculty Characteristics and Perceptions

The literature supports that faculty are divided on many issues surrounding the SET: usability, reliability, and validity. Although some faculty find ratings useful, SET are not largely perceived as being reliable or valid (Balam & Shannon, 2010). SET are by no means universally useful to faculty and are sometimes viewed by faculty more as a measure of popularity than of effective teaching (Beran & Rokosh, 2009a, 2009b; Beran et al., 2005). Faculty agree that students are not pedagogical experts (Ackerman et al., 2009). El Hassan (2009) supports that faculty could be more supportive of SET, if only they were not significantly incorporated in promotion and tenure decisions.

The portion of the literature that reviewed personal traits of faculty that rate the highest in the eyes of students are universally desirable characteristics that faculty would likely agree upon: experience, clarity, enthusiasm, helpfulness, fairness, and organization (Barth, 2008; Donnon et al., 2010; Nowell et al., 2010; Sitzman, 2010; Wolf et al., 2004). There is a gap in the research as to whether student definitions are identical to faculty definitions of these characteristics. For example, faculty definition of clarity might differ from a student definition; timeliness, a component of organization, possibly differs generationally between faculty and student.

Three nursing-specific studies mentioned the importance of faculty-caring capabilities. This involves extending the nurse-to-client caring philosophy into the caring that an instructor shows for his or her students (Sitzman, 2010; Tang et al., 2005; Wade & Kasper, 2006). Exactly what makes up these caring behaviors needs further examination, however. It is unclear if students in other disciplines single out caring behaviors as important. It is possible that caring behaviors are labeled under a different name in other disciplines because of philosophy differences. For example, in a study of business students, a willingness of faculty to help students has been identified as key in overcoming negative SET in more rigorous classes (Hills et al., 2009). What constitutes caring in nursing education is likely an important area to address in future instrument development and also in considering how nursing instruction might be different than other disciplines.

Student Characteristics and Perceptions

Research related to the theme of student characteristics is important because it helps faculty compensate for these characteristics, which are generally beyond faculty control, and tend to lead to lower faculty ratings in SET. These factors include student age (younger students are more critical in SET), expected grade (a lower expected grade tends to lead to a lower SET), student engagement and attendance (positively correlated with ratings), and full-time versus part-time status (part-time students rate higher; Arreola, 2007; Barth, 2008; Beitz & Wieland, 2005; Beran & Violato, 2009; Serdyukova et al., 2010).

Specific program type within nursing instruction was not significantly correlated to differences in SET, but this area is underresearched (Beitz & Wieland, 2005). These student characteristics are not to be used as excuses by faculty to explain low ratings to an administrator but rather as a baseline contextual picture of student-related factors that influence SET. In addition, students, like
faculty, do not necessarily view the SET process favorably. This can sometimes lead to poor execution of SET and potentially unfavorable evaluations. For example, students sometimes will indicate a neutral (3) rating, even when an issue is nonapplicable (Emerson & Records, 2007) or admit to making false statements on SET (Clayson & Haley, 2011). This is an aspect of SET that departments should consider when reviewing policies related to SET.

Course Characteristics
Knowing how course characteristics affect SET can assist those trying to place the SET in context and maximize its usefulness. For example, SET scores vary by factors such as if a course is mandatory or not, what kind of room a course is held in, and if a class is large (Beran & Violato, 2009; Hill & Epps, 2010). Not surprisingly, high workload negatively affects SET (Beran & Violato, 2005; Donnon et al., 2010). Relevant to nursing is the finding that laboratory classes seem to receive higher ratings than lecture courses (Beran & Violato, 2005); however, these results were only cited in one study. Unknown is how practica might be rated differently by students, if at all. The important piece to glean is that class format might affect ratings.

On-line Versus Paper SET
There was limited data on how an on-line SET is compared with paper-and-pencil SET. On-line, out-of-class SET lowers faculty ratings and likely leads to polarized evaluations and a low response rate (Nowell et al., 2010; Serdyukova et al., 2010). Conflicting data were found regarding the effect that on-line SET have on the resulting numbers. Because many programs are increasingly adopting an exclusively on-line format or a combination thereof, this is an area for future research.

Reliability and Validity of SET
Generally, reliability and validity of student ratings are supported (Appling, Naumann, & Berk, 2001; Arreola, 2007). Although students are not content experts, they can serve as reliable and valid evaluators of delivery method and teaching effectiveness. They can rate the response to a class or can compare with other teaching styles to evaluate what worked best for them (Ackerman et al., 2009). Several instruments in the nursing literature have been tested and have been found to be valid and reliable (Emerson & Records, 2007; Tang et al., 2005).

When viewed in proper context, SETs are an invaluable picture into faculty performance. Even a good instrument is still subject to the variations cited in the literature (faculty, student, and course characteristics). However, knowing how these variations play into SET is integral to their interpretation, particularly when administrators use SET to compare faculty or to argue for or against tenure and promotion. Used properly, SET can be important tools for faculty development (Arreola, 2007).

Development of New Instruments
The key to an effective SET is a good instrument, one that faculty can trust. It must be tested, valid, and reliable for it to be taken seriously. A good instrument recognizes faculty strengths, yet provides direction for faculty growth, and reduces or eliminates inappropriate student comments. Many universities use homegrown scales, do not subject these instruments to rigorous testing, and do not reevaluate them over time, thus never doing what is required to create a quality instrument (Arreola, 2007; Hills et al., 2009). Faculty know this, and this knowledge contributes to mistrust and underutilization of SET.

As a result of this literature review, it is clear that unique characteristics within nursing necessitate a very specific SET. Two nursing researcher teams have developed instruments that incorporate a wide range of areas of importance to nursing and a wide range of class type differences (Emerson & Records, 2007; Tang et al., 2005). For instance, Emerson and Record's (2007) use of five domains (communication, expertise, pedagogy, advocacy, and professionalism) or Tang et al.'s (2005) use of four domains (professional competence, interpersonal relationships, personality characteristics, and teaching ability) are appealing because they incorporate items that carry across all types of class formats and address a variety of faculty characteristics (Emerson & Records, 2007; Tang et al., 2005). Nursing faculty should incorporate work of previous researchers because they consider future instrument development.

At many institutions, administration controls the choice of instrumentation. The literature indicates that administrators weigh the item addressing the overall quality of course instruction as the most useful factor. The impetus for instrument development or revision in these studies was precipitated by faculty concerns. Resistance from administration was not addressed in any of the articles. Of SET, faculty rating instruments, open-ended comment forms, unit-specific rating instruments or teaching portfolios, and student evaluations are rated as most important by administrators when evaluating teaching for merit pay, promotion and tenure, hiring, and retention (Beran et al., 2005). Beran et al. (2007) also emphasize that most of the administrators have been instructors and may be aware of the implications of student evaluations. In addition, SETs may be mandated by institutional policy or by student demand, and the instrument may be institution wide versus unique to the teaching unit. Both administrative control and institutional policies can be potential barriers to new instrument development.

Limitations
The main limitation of the literature review is the small amount of nursing-specific research. There is little research contrasting different types of programs within nursing, and scarce information of varied class formats within nursing. However, strength of the literature review is the broad nature of the research reviewed. This creates confidence in the generalizability of findings.
Implications for Practice

The problem lies not in arguing over whether SET are reliable or valid or not, or comparing various characteristics, but rather in determining what is the best way to incorporate SET, viewed in context, to improve teaching and learning. Triangulation is one method that has been proposed to promote a more solid, varied measure of faculty teaching effectiveness from the perspectives of all involved: faculty, students, and self (Appling et al., 2001). SET are incredibly useful as one part of a three-pronged system of faculty effectiveness. Faculty should be evaluated by peers in the department or, at the very least, observed and mentored (Appling et al., 2001). Students are not content area experts, nor faculty delivery effectiveness experts (Ackerman et al., 2009). The teaching portfolio (self-evaluation) becomes an important aspect in a triangulated teaching evaluation and is a way for the instructor to “prove” their worth as a creative educator by exhibiting syllabi, describing one’s philosophy, and discussing individual interpretations of SET and peer-teaching evaluations. Using a triangulated system helps compensate for the shortcomings of other methods and, when used together, can stand as a good measure of teaching effectiveness.

It is clear that teaching takes place in a highly contextual and variable environment. SET can be important indicators of what would help students to learn, provided that they are asked the right questions (Richardson, 2005). If instruments are well designed, they will be more usable. This literature review not only provides the context in which to interpret SET but also supports an expanded, more complete view of faculty evaluation—one that can more accurately discriminate who is a truly expert instructor and to assist those who are less than expert with concrete and useful ways to develop professionally.

References


