Florida Journal of Educational Research
Fall 2001, Vol 41, No 1, 14-42

A Series Of Studies Examining The Florida Board Of Regents' Course Evaluation Instrument

Tary L. Wallace

University of Central Florida

Lynn Grinnell
Lou M. Carey
Robert F. Dedrick
John M. Ferron
Kathleen A. Dailey
Dorian Vizcain
James A.White

University of South Florida

Abstract

This research examined the psychometric properties (e.g., factor structure, reliability) of the Florida Board of Regents Student Assessment of Instruction instrument and the relation between various factors (adaptations for distance education, initial expectations, time, non-instructional factors, and response scale format) and students' course evaluations. Data were collected from 631 students in an undergraduate course in educational assessment and in graduate courses in educational technology, language arts, and library science at various times during the semester. Results for the course evaluations reflected a one-factor model and internal consistency reliabilities greater than .90. No significant differences in students' course evaluation ratings emerged across time during the semester, students' first and last day ratings of a course, non-instructional factor,(excluding hours employed), or response scale formats.

The Board of Regents (BOR) of the State University System of Florida mandated that each state university in Florida use the State University System Student Assessment of Instruction (SUSSAI) instrument beginning in the spring of 1996. With limited exception, all undergraduate and graduate courses taught by faculty members, adjuncts, and graduate assistants were to be assessed using this instrument. It was also mandated that summary results be made available to students or members of the public to facilitate student selection of courses and that results be used in the evaluation of faculty instruction (State University System of Florida, 1995). The mandated introductory statement and eight items may be supplemented with other assessment items used by a university, college, or department.

Best practices would suggest that psychometric properties of the instrument be examined before results are used for making meaningful decisions (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999).

Though the Board of Regents of the State University System of Florida has since been dissolved by the governor, it is anticipated that current policies will remain in effect at least until the new management system is in place; continuation thereafter remains to be determined. Since these changes in the university management system are imminent, research to inform decisions regarding the continuation, revision, or elimination of current practices related to course evaluation is needed.

The purpose of the article is to examine the Florida Board of Regents Student Assessment of Instruction (1995) instrument and to invite policy analysis discourse on the evaluation of university instruction. The six studies examine psychometric properties of the instrument and the relation between various factors and students' ratings of their professors as measured using the SUSSAI professor rating form. The studies are as follows: (1) Factor Structure of the Florida SUSSAI Professor Rating Form, (2) Using a Modified Board of Regents Course Evaluation Form for Distance and Technology-based Courses, (3) Comparison of Students' Initial Expectations For Their Professors and Their End of Course Ratings, (4) Changes in Individual Students' Evaluations of Their Professors Over a Semester, (5) Non-Faculty Factors Related to Students' End of Course Ratings of Their Professors, and (6) Influence of Response Scale Format on Students' Ratings of Professors.

Theoretical Background

Students' Ratings of Instruction

Use of student ratings of faculty instruction has grown notably since the 1960's to the extent that it is nearly a universal practice (Centra, 1993). Typically, questionnaires are used to collect student perceptions of the course and instructor, with the information being used for (a) diagnostic feedback for faculty, (b) measures of effectiveness for administrative decision making, (c) students' selection of courses, (d) course or curriculum development, and (e) research on teaching (Marsh, 1987). The quality of information from such assessment is related to the psychometric properties of the instrument, particularly the validity and reliability of the obtained results. The most widely used instruments are similar in content, and their results tend to have reasonable estimates of validity and reliability (Centra, 1993; Marsh, 1987). Researchers have examined the relationships between student ratings and various indicators of effective instruction (e.g., student achievement, instructor self-ratings, administrator and colleague ratings) and have concluded that student ratings do correlate positively with these indicators (Cashin, 1994). Ratings are considered generalizable across a number of instructional situations and unaffected by bias, though this issue continues to be examined in different contexts and related to specific sources of bias (Feldman, 1997; Greenwald, 1997; Marsh, 1987). Prior research has indicated students' ratings of instruction correlate with various background characteristics such as prior interest, general interest, workload, and expected grade (Cashin, 1994; Marsh, 1987).

Since no single definition of effective instruction is universally accepted, there are many similar instruments that differ according to each author's theorized model of the construct. Effective instruction is generally thought to be multidimensional; however, researchers differ in their definitions of the dimensions and the role of dimensionality within the various purposes of instructional evaluation (d'Appollonia & Abrami, 1997; Marsh & Roche, 1997). Although student ratings of instruction have been studied extensively, it is generally agreed that there is a need to study specific instruments for the particular purposes for which they are used (Arreola & Aleamoni, 1990; Dilts, Haber, & Bialik, 1994; McKeachie, 1997).

Samples and Procedures

Participants for the six studies included pre-service teacher education students enrolled in an undergraduate assessment course and graduate, web-based distance education students enrolled in educational technology, language arts, and library science courses (N = 631) at a large state university in Florida. The preservice teacher sample was drawn from 26 sections of the assessment course taught by four instructors from the fall 1995 through spring 1999 semesters. Students' majors included elementary (43.5%), secondary (19.5%), exceptional student education (20.5%), performance areas such as art, music and physical education (12%), and other (4.5%). A large majority of the students were female (80%) as is typical of pre-service teacher education programs. Students registered for the assessment course following typical course registration procedures, and without knowing in advance the instructor assigned to their section. Two instructors were full time professors and two were adjunct instructors who were doctoral students in the subject with more than three years college level teaching experience. Students enrolled in the graduate educational technology, language arts, and library science courses were majors in these program areas, and they had advance knowledge of the distance format and the faculty who would teach their courses.

For the assessment course, each section used the same syllabus, instructional equipment and materials, as well as the same assessments. The course was comprised of four units of instruction. Students completed a portfolio and each unit was followed by a criterion-referenced, objective, unit posttest and two attitude questionnaires. Achievement and attitudinal pre-assessments were administered the first day of class during course orientation but prior to any discussion of the course. The final attitude questionnaires were completed following the final exam but prior to assignment of course grades. The graduate courses all differed by content, instructional procedures, and faculty.

Instruments

State University System Student Assessment of Instruction (SUSSAI). This instrument, developed by a task force in accordance with a mandate by the Florida State Board of Regents, is used to collect and publish students' ratings of faculty and instruction. It was influenced in content and format by an instrument previously used at a large state university. A faculty committee at that university conducted a literature review and compared instruments from universities throughout the United States. The criteria used to design the items were that the questions would be general and applicable in various settings, not diagnostic in nature, and used for general rather that specific interpretation. After the committee agreed on content and format, a pilot was conducted and factor analysis used to examine data from the newly formed general questions and a set of special questions on testing and grading, course organization, assignments, and amount learned (Legg & Cunningham, 1995). The BOR task force, working through memos and phone conversations, developed a draft of its version that was circulated among the task force and various constituencies. The existing SUSSAI content and format was then finalized (J. Eison & J. Linder, personal communication, October 1997). The SUSSAI is an evaluation form that solicits student responses to eight global statements about instructional content and delivery (see Table 1).

Academic Motivation Profile (AMP). The Academic Motivation Profile (Carey, 1990) is based on theories of students' academic motivation described in Keller's ARCS Model (1987a, 1987b) of instructional design. Keller's model includes four elements that are critical to the design of effective instruction: attention (A), relevance (R), confidence (C), and satisfaction (S). The AMP provides a measure of students' evaluation of a course relative to their attentiveness during instruction; the relevance of the material presented for their personal and professional goals; the level of confidence they develop during the course in performing course goals; and intrinsic satisfaction with their own participation, development, and professional affiliation. It contains four factors with nine items each. Previous studies (Carey, Dedrick, Carey, & Kushner, 1994) of the psychometric properties of the AMP supported the four-factor structure with high estimates of Cronbach alpha internal consistency reliability indices for the subscales: attention, .90; relevance, .91; confidence, .95; and satisfaction, .92.

Instrument Administration Procedures

For the assessment course, students completed the SUSSAI and the AMP during orientation on the first day of class prior to any discussion about the course or faculty expectations and again following each of four achievement tests, including the final examination on the last day of the course. Students' anonymity was protected, and to link students' responses across time, they used a code consisting of their fathers' and mothers' first names. The undergraduate and graduate students in the web-based distance courses completed the SUSSAI only once during the last week of class.

The two instruments used in the studies with assessment students were direct adaptations of the SUSSAI and the AMP. They were administered together in one of three forms, each using the same items as the original instruments but a different response format (Form A, B, or C). Form A contained the original SUSSAI response scale which ranged from 1 Excellent to 5 Poor; Form B contained the current scale which ranged from 1 Poor to 5 Excellent; and Form C contained a Likert-type scale that ranged from 1 Strongly Disagree to 5 Strongly Agree. Item stems for the SUSSAI and the AMP remained the same across all forms. Prior to initial data collection, students were randomly assigned to one of the forms that they continued to use during the four remaining data collection points throughout the semester.

The SUSSAI was adapted for web-based distance students to reflect better the instructional context in distance education. Only two items were modified, and prior to administration, the item adaptations were judged appropriate by the University's Academic Computing Committee, Instructional Technology Committee, Distance Education Learning Model Work Group, and the College of Education's Instructional Technology Committee. The item adaptations are included in Table 1. The remaining sections of this paper describe the six studies examining the characteristics of the SUSSAI.

Table 1
State University System Student Assessment of Instruction Items and Items Adapted for Distance Education

Items	Responses[a]

	P	F	G	VG	E
1. Description of course objectives and assignments	1	2	3	4	5
2. Communication of ideas and information	1	2	3	4	5
3. Expression of expectations for performance in this class	1	2	3	4	5
4. Availability to assist students in or out of class (Availability to assist students)	1	2	3	4	5
5. Respect and concern for students	1	2	3	4	5
6. Stimulation of interest in the course	1	2	3	4	5
7. Facilitation of learning.	1	2	3	4	5
8. Overall assessment of instructor (Overall assessment of instruction)	1	2	3	4	5

Note. Items in parentheses were adapted for distance education.
[a]P = Poor, F = Fair, G = Good, VG = Very Good, and E = Excellent.

Study 1: Factor Structure of the Florida SUSSAI Professor Rating Form

Confirmatory factor analysis is frequently used to determine the extent to which data fit a theoretical model from which an instrument has been designed thereby providing some of the necessary evidence of validity (Crocker & Algina, 1986). The procedure is frequently used to determine the extent to which data from students' evaluations of teaching effectiveness fit dimensions identified in models of effective instruction (Marsh, 1987). Logical analysis of research on instructional practices, student achievement, and other, similar instruments often guide the development of these models. Items are then written to represent various aspects of effective instruction within the model. Data are collected and factor analysis is used to determine whether students' responses to the questions fit the hypothesized model. There is similarity among the models but researchers have differed on the number as well as the nature of dimensions identified and how they relate to use of the ratings. Marsh (1987) has proposed nine dimensions of effective instruction, Feldman (1997) proposed 28, and Centra (2000) proposed eight dimensions. It is generally agreed that scores from multidimensional instruments are more useful for formative evaluation of instruction. There is less agreement about whether more specific scores from multidimensional instruments or general information from global items would best serve personnel evaluation purposes (d'Apollonia & Abrami, 1997; Marsh, 1997). Other studies have examined whether the many aspects of instruction may be represented with two or three higher order factors such as pedagogical skill and rapport; professional maturity and empathy; or presentation, facilitation of learning, and management (see

Marsh, 1991) or even a single, global factor (Cheung, 2000; Marsh, 1991). Best practices suggest that developers and users of such instruments provide evidence that inferences made from the scores will be appropriate for the specified purpose.

The SUSSAI has been developed to assess instruction for accountability purposes. It will be examined here in relation to a three dimensional model of effective instructional practices developed from cognitive learning theory and research on the effects of interpersonal interactions between students and teachers on classroom learning. The proposed instrument dimensions are generated from Gagné's nine instructional events (Gagné, 1985). The first dimension, Student Preparation, concerns preparation of the learner for instruction and reflects Gagné's gaining attention, informing the learner of the objective, and stimulating recall of prerequisite learning events. The second dimension, Teacher Preparation, consists of events related to processes involved in the delivery of instruction: presenting the stimulus material, providing learning guidance, eliciting the performance, and enhancing retention and transfer. Using a more global perspective, these two dimensions may be viewed as part of a single construct that includes the pedagogical aspects as they are distinguished from the personal, social aspects of instruction. The last dimension, Interaction, is derived from Gagné's providing feedback and assessing performance events as well as research on interpersonal aspects of teacher/student interactions that have shown positive relationships to learning. The interaction dimension also includes the social context of learning. Tiberius and Billson (1991) believe that this context includes, but is not limited to, the roles, responsibilities, and interactions between teachers and students that are present in every instructional situation. They refer to an alliance between teachers and students that promotes student growth. Communication theory and the notion that communication between parties is influenced by the nature of the relationship between them are central in their discussion of social context. Two of the seven principles for good practice described by Chickering and Gamson (1991), including student faculty contact and respect for diverse talents and ways of learning, are interaction related. This interpersonal dimension resembles empathy or rapport from the more global perspective.

This study used confirmatory factor analysis to evaluate the internal structure of the SUSSAI instrument. Content analysis yielded items that logically fit one of the dimensions described above rather than the others. Because the instrument contains only eight items, seven that imply specific behaviors and one global item, and because of the limitations of confirmatory factor analysis with such a small number of items, the hypothesized dimensions were represented in the model as follows. The teacher preparation and student preparation dimensions were combined to form an instructional skill dimension that would be distinguishable from the interaction dimension. Two alternative factor models were tested: (a) a two-factor model in which five items (Description of course objectives and assignments, Communication of ideas and information, Expression of expectations for performance in this class, Stimulation of interest in course, Facilitation of learning) loaded on an Instructional Skill factor, and two items (Availability to assist students in or out of class, Respect and concern for students) loaded on an Interpersonal factor, and (b) a one-factor model in which seven-items loaded on one factor (item 8 "Overall assessment of instructor" was not used for these analyses because it represented a summary measure).

Sample and Procedures

Undergraduate students (n = 495) from multiple sections of a course in classroom assessment participated in the study between fall, 1995 and spring, 1997. Students were randomly assigned to form A, B, or C of the course evaluation questionnaire, as described previously. Students anonymously and voluntarily completed the instrument following their final exam and prior to receiving their term grade.

Instruments and Analyses

The State University System Student Assessment of Instruction (1995) instrument described previously was used in this study. Confirmatory factor analysis models were estimated using polychoric correlations and the weighted least squares fitting function in LISREL 8 (Joreskog & Sorbom, 1995). One- and two-factor models were evaluated for each form of the SUSSAI instrument (A, B, C).

Results and Discussion

The fit statistics for the one-factor model, while not as good as those for the two-factor model, indicated that the model provided an acceptable fit to the data. Additional support for the one-factor model is suggested by the strong correlation between the Skill and Interpersonal factors in the two-factor model for Forms A, B, and C (rs = .95, .90, .92). These correlations indicate that there is considerable overlap between the Skill and Interpersonal factors. Therefore, the one-factor model provides an acceptable and parsimonious representation of the data. Table 2 summarizes the fit statistics for all the models.

Study 2: Using a Modified Board of Regents Course Evaluation Form for Distance and Technology-based Courses

The SUSSAI form was reworded to make it more appropriate for the instructional context in distance and other technology-based courses. The purpose of this study was to examine the impact of the reworded items on the factor structure of the overall instrument.

Sample and Procedures

The participants in this study included 99 undergraduate and graduate students enrolled in web-based distance courses in assessment, educational technology, language arts, and library science at the university. The instrument was administered only once in the spring term of 1999 during the last week of class; students completed the form anonymously. There were 91 cases with complete data.

Instruments and Analysis

Two modifications were made to the wording of the SUSSAI items. One item originally asked students to rate the instructor’s “availability to assist students in and out of class.” The item was truncated to instructor’s “availability to assist students,” making the item less specific to classroom instruction. The last item originally asked students to rate the “overall assessment of instructor.” The item was modified to rate the “overall assessment of instruction,” making the item less specific to the one-instructor classroom model and more applicable to distance and technology courses developed and delivered by teams (see Table 1). Question 8, “Overall assessment of the instructor,” was excluded from the factor analysis in Study 1, since it is a summary item. With the modification of Question 8 to read, “Overall assessment of the instruction,” it potentially fit with the Instructional Skill factor. As such, it was included in the factor analysis for Study 2. Confirmatory factor analysis was used to examine the factor structure of the modified instrument, and results of this analysis were then compared to those of Study 1.

Results and Discussion

A summary of the fit statistics for the two models tested in Study 2 is in Tables 2 and 3.

Table 2

Fit Statistics for One- and Two-Factor Models for the Board of Regents Course Evaluation Instrument by Form (Study 1)

Study 1 (n = 495)						
	One factor			Two factor		
Fit statistic	Form A	Form B	Form C	Form A	Form B	Form C
?^{2}	25.778	36.865**	31.411*	19.061	27.013	17.099
Df	14	14	14	13	13	13
RMSEA[a]	.071	.101	.087	.053	.082	.044
CFI[b]	.998	.998	.995	.999	.999	.999

Note. Form A: n = 168, Form B: n = 160, Form C: n = 167.

[a] *RMSEA = Root mean square error of approximation (RMSEA < .08 represents acceptable fit).*

[b] *CFI = Comparative fit index (CFI > .90 represents acceptable fit).*

* *p < .01. ** p < .001.*

Table 3

Fit Statistics for One- and Two-Factor Models for the Instrument Modified for Distance and Technology-based Courses (Study 2)

Study 2 (n = 91)		
	One factor	Two factor
?^{2}	29.02	27.45
df	20	19
RMSEA[a]	0.07	0.07
CFI[b]	1.00	1.00

[a] *RMSEA = Root mean square error of approximation (RMSEA < .08 represents acceptable fit).*

[b] *CFI = Comparative fit index (CFI > .90 represents acceptable fit).*

* *p < .01. ** p < .001.*

Results in Study 2 using the modified SUSSAI are very similar to those of Study 1 using the current form of the instrument. Again, the one-factor model provided an acceptable and parsimonious representation of the data. A two-factor model also fit, with one factor representing "Instructional skills" and the other factor "Interpersonal skills." As stated in Study 1, support for the one-factor model is strengthened by the strong correlation between the Skill and Interpersonal factors in the two-factor model, indicating considerable overlap. The same is true for the modified items. The modest wording changes to the SUSSAI to make it more applicable to distance and technology-based courses did not change the factor structure of the instrument. The advantage of using the modified SUSSAI with students in distance courses is that it poses questions that are more applicable to students in the rapidly growing number of distance and technology-supported courses at universities.

Study 3: Comparison of Students' Initial Expectations For Their

Professors and Their End of Course Ratings

Although pretesting achievement for course evaluation is considered prudent, the practice has not generalized in most cases to also pretesting students' attitudes about a course. Students entering a course possibly have expectations for the course and instructor that affect their attention to various course details and their interpretation of events that occur during the class. Applying the value-added concept to students’ course ratings and interpreting post-course evaluations in light of pre-course expectations should enable more valid inferences about course quality than do single post-course evaluations.

The validity of using students’ end of term evaluations to reflect course quality and individual instructor effectiveness has been questioned for years. Crittendon and Norr (1973) cautioned that student end of term evaluations should not be used to infer instructional effectiveness since these ratings include an interaction between student values and teacher behaviors. In the same vein, Finaly and Neumann (1985) described students' satisfaction with instruction as a "generalized attitude toward various facets of college life which is cumulatively influenced by students' experience with instructors in college" (p. 11); these generalized attitudes illuminate more about the college as an organization than about the individual instructor of a course. In spite of these cautions, end of term student evaluations are commonly recommended as legitimate outcome measures for accreditation purposes, and they are used by faculty and administrators to infer instructional effectiveness.

These values or generalized attitudes are related to students' expectations. Their expectations in a given setting affect how they perceive the event or experience, and their differing expectations help to explain how different students give divergent accounts of the same instructor and course. Slavin (1991) suggests that individuals' perceptions of experiences are influenced by their mental state, past experiences, knowledge, motivations, and other factors. Students enter a classroom with expectations for their progress and instructor based on their history of success and failure in school, the reputation of the course, and the reputation of the instructor (Kaplan, 1990). Their expectations have been observed to affect many aspects of their behavior including attention, persistence, and effort.

Slavin (1991) describes two types of motivation: (a) generalized motivation, a relatively stable personality characteristic that is shaped by an individual's history of reinforcement and (b) situation motivation, a situation-specific attitude that results from a particular circumstance. He indicates that while situation motivation changes with events, generalized motivations "tend to remain constant across a variety of settings and are difficult to change in the short run" (p. 329). Generalized motivation helps to explain students' general attitudes about school and instruction while situation motivation helps to explain changes in students' attitudes relative to a particular course and instructor. The problem with interpreting students' end of term ratings of a course and instructor is determining whether they mostly reflect their generalized motivations about school or whether they mostly reflect situation motivations related to the particular instructor and course. The purpose of this study was to compare students' first day and last day ratings using the SUSSAI.

In a prior study (Carey, Carey & Pearson, 1992), students' first day and last day evaluations of a course were compared using the College of Education's Student Evaluation of Teaching (SET). The sample included 199 students enrolled in multiple sections of the undergraduate assessment course. ANCOVA for the last day evaluation revealed that students first-day expectations were the best predictors of last day evaluations with significant predictions observed for day one expectations (p < .0001), class section (p < .0001), and section-by-major interaction (p < .0148). ANOVA for the last day SET only indicated significantly different course evaluations by class section (p < .0001), no significant differences by major (p < .9604), but a significantly different section-by-major interaction (p < .0229). Similarly, ANOVA for the day one SET measure illustrated that students' expectations for the instructor also differed significantly at the outset for class section (p < .0045) and major (p < .05), but not for section-by-major interaction.

Sample and Procedures

The sample (N = 152) included only those students who completed the instruments on both the first and last day of class during Fall of 1996. Students were randomly assigned to form A, form B, or form C of the SUSSAI. They completed the forms on the first day of class prior to any course orientation or instruction and again on the last day of class following the final examination but prior to learning their term grade. Individual students' pre-post SUSSAI forms were linked using parents' first name codes known only to them. No significant differences in student responses were observed among the three response formats; therefore, the data were combined for the three forms, with a scale of 1 (poor) to 5 (excellent).

Descriptive statistics were calculated for item and overall instrument means and standard deviations and to check the assumption of normality. Since SUSSAI data are aggregated and reported by item means, item means collected on the first and last day of the course were compared, using paired-samples t tests.

Results and Discussion

Based on Slavin's (1991) description of generalized motivation, a relatively stable personality characteristic that is shaped by an individual's history of reinforcement; prior research with the SET; and the global nature of the eight items on the SUSSAI; no significant differences were expected between students' first day judgments (expectations) and last day ratings of the instructor and course using the SUSSAI.

Table 4 contains the means and standard deviations for the eight items on the SUSSAI collected on the first and last day of class following the achievement posttest. Items were rated on a scale from 1 (poor) to 5 (excellent). Examining these end of course data in the typical manner they are reviewed by personnel committees, administrators, and the general public in the university libraries across the state, the reader might look for (a) the general position of each item on the five-point scale as in indicator of overall course quality, (b) areas where student ratings differ using the comparative sizes of the standard deviations across the items, and (c) relative strengths and problems, defined as areas with most room for improvement across the items, based on highest and lowest ranked items in the set.

Based on the item means for the last day of class in Table 4, it appears that students considered the course to be "pretty good" overall since all item means were greater than 3.5, and the item mean average was 3.82 (SD = 0.96). Examining "agreement" among students using the standard deviation, there was most agreement on item 1, description of course objectives and assignments (SD = 1.00), and least agreement on item 5, respect and concern for students (SD = 1.19). The differences observed in agreement are possibly related to the relative abstractness of the two items. Considering relative strengths and problems, it appears that the course strengths are item 8, overall assessment of instructor (M = 4.05, SD = 1.11) and item 4, availability to assist students in or out of class (M = 4.00, SD = 1.00). Areas where the course has most room for improvement include item 6, stimulation of interest in the course (M = 3.52, SD = 1.18) and item 2, communication of ideas and information (M = 3.70, SD = 1.03).

Table 4
Item Means and Standard Deviations for the SUSSAI on the First and Last Days of the Semester and t-tests Comparisons Between Occasions

Items	First day (pre)		Last day (post)		Difference			
	M	SD	M	SD	M	SD	t	p
1. Description of course objectives and assignments	3.89	.87	3.86	1.00	-0.04	1.17	-0.42	.677
2. Communication of ideas and information	3.83	.87	3.70	1.03	-0.14	1.22	-1.40	.164
3. Expression of expectations for performance in this class	3.88	.91	3.82	1.03	-0.06	1.29	-0.57	.571
4. Availability to assist students in or out of class	3.83	.87	4.00	1.09	0.17	1.30	1.62	.107
5. Respect and concern for students	4.03	.93	3.88	1.19	-0.15	1.41	-1.32	.187
6. Stimulation of interest in the course	3.70	1.03	3.52	1.18	-0.18	1.29	-1.76	.081
7. Facilitation of learning	3.74	.84	3.75	1.05	0.01	1.20	0.14	.892
8. Overall assessment of instructor	4.13	.86	4.05	1.11	-0.07	1.25	-0.71	.476
Mean of all items on instrument	***3.88***	***.73***	***3.82***	***.96***	***-0.06***	***1.03***	***-0.69***	***.491***

Note. n = 152.

To enable a comparison between students' ratings of faculty on the last day of the semester with the same students' ratings on the first day of the term, Table 4 also includes the item means and standard deviations for both occasions as well as the mean differences, t values, and p values for each item. Comparisons were made using paired-samples t tests, and no significant differences were observed on any one of the eight items or on an instrument composite score between students' ratings of the instructor on the first and the last day of the course. The t values ranged from $t = –1.76$, ($p < .08$) to $t = 0.14$, ($p < .89$).

When student rating data are aggregated by item means, as they are for reporting results to faculty members, administrators, and the public, there are no significant differences between students' day one and last day ratings on an item-by-item or overall instrument basis for this sample. These results are consistent with Slavin's generalized motivation rather than his situation motivation theory, a situation-specific attitude that results from a particular circumstance. As noted previously, while situation motivation changes with events (a particular course and particular faculty member), generalized motivations "tend to remain constant across a variety of settings and are difficult to change in the short run (p. 329)."

This causes one to question the validity of the manner in which the SUSSAI is used and interpreted as a measure of individual faculty effectiveness. On this issue, the reader should pay attention to the changes in students' attitudes measured using the SUSSAI in the following study since the results are consistent with this study, and a much larger sample of students is used.

Study 4: Changes in Individual Students' Evaluations of Their Professors Over a Semester

Much course evaluation research has involved collecting data at the end of the course (i.e., one-wave) or at midsemester and the end of the course (i.e., two-wave). Although these data collection strategies provide formative and summative feedback to faculty, they provide limited information about how students' perceptions of instruction change during the course of a semester. In a prior study using the AMP (Dedrick et al., 1995), a five-wave design (day 1 of class, and weeks 4, 8, 11, and 15) was used to measure students' perceptions of four dimensions (Attention, Relevance, Confidence, and Satisfaction) of a course. Results of this study showed that students became less interested in the course and viewed the course as less relevant over time. No changes were observed in students' levels of confidence and satisfaction. To extend this research, the present study used growth curve analysis (Bryk & Raudenbush, 1992) to examine the change in the overall score of the SUSSAI course evaluation instrument at four points during the semester.

Sample and Procedures

Undergraduate students (n = 165) enrolled in six sections of an introductory assessments course participated in the study during the fall semester, 1996. Within each of the six sections of the course, students were randomly assigned to one of three response scale formats (Forms A, B, and C). Comparisons of the demographic characteristics of the students across the three formats revealed no significant differences (p's > .05) on number of semester hours completed, average number of hours employed, experience with math and computers, and interest in teaching.

Instruments and Analysis

Students completed the SUSSAI and the AMP five times during the semester (day 1 of class, and weeks 4, 8, 11, and 15). Internal consistency reliabilities (Cronbach alpha) of the measures were all greater than .89 (range for the SUSSAI was .92 at week 4 to .97 at week 15; for the AMP the range was .89 for Attention at week 4 to .97 for Confidence at week 15). Analyses for the present study were based on the data collected at four time points (weeks 4, 8, 11, and 15).

Results and Discussion

Table 5 includes the overall SUSSAI scores (mean of the 8 items) and the individual items over the four time points. The means for the overall SUSSAI score ranged from 3.54 (SD = 0.98) at Week 8 to 3.81 (SD = 0.98) at Week 15. Linear growth trajectories derived using HLM/2L (Version 4.01; Bryk, Raudenbush, & Congdon, 1994) for the SUSSAI measure indicated that on average the students' evaluations did not change significantly over the course of the semester (slope = 0.004, p = .526). Analyses conducted separately on each response scale format revealed a similar pattern of non-significant changes in students' evaluations. The slopes for Form A (original scale), Form B (reversed scale), and Form C (Likert agreement scale) were 0.004, 0.009, and -0.008, respectively. These slopes were not significantly different from zero and were not significantly different from each other.

In contrast to the non-significant changes observed on the SUSSAI Form, significant changes were observed for these same students on three of the subscales from the AMP (see Table 6). Students' perceptions of the relevance of the course declined over the semester (slope = -0.025, *p* < .001), while their sense of confidence in performing the skills learned in class and their satisfaction with their personal development increased during the semester (slopes = 0.028 and 0.018, *p* < .001 and *p* < .01, respectively). The fact that changes were observed on the AMP and not the SUSSAI form suggests either that the SUSSAI instrument may not be sensitive enough to detect changes occurring during the course of the semester or that the construct measured by the SUSSAI instrument is relatively stable over time.

Table 5

Descriptive Statistics for the SUSSAI Course Evaluation Form by Week

Variable		Week 4	Week 8	Week 11	Week 15
SUSSAI Overall	M (SD)	3.69 (0.82)	3.54 (0.98)	3.58 (0.91)	3.81 (0.97)
1. Description of course objectives and assignments	M (SD)	3.83 (1.02)	3.58 (1.08)	3.69 (1.01)	3.88 (0.99)
2. Communication of ideas and information	M (SD)	3.38 (1.08)	3.16 (1.15)	3.38 (1.04)	3.72 (1.02)
3. Expression of expectations for performance in this class	M (SD)	3.85 (0.95)	3.68 (1.05)	3.69 (0.99)	3.85 (1.01)
4. Availability to assist students in or out of class	M (SD)	4.08 (0.91)	3.90 (1.05)	3.86 (1.11)	4.03 (1.08)
5. Respect and concern for students	M (SD)	4.11 (0.96)	3.86 (1.16)	3.82 (1.16)	3.91 (1.18)
6. Stimulation of interest in the course	M (SD)	3.25 (1.19)	3.19 (1.21)	3.22 (1.20)	3.53 (1.16)
7. Facilitation of learning.	M (SD)	3.36 (1.02)	3.41 (1.12)	3.40 (1.00)	3.79 (1.04)
8. Overall assessment of instructor	M (SD)	3.68 (0.97)	3.51 (1.18)	3.55 (1.03)	3.79 (1.09)

Table 6

Intercepts and Slopes for Linear Growth Curves for Board of Regents (SUSSAI) Course Evaluation Form and Academic Motivation Profile (AMP)

Variable	Intercept	Slope
SUSSAI	3.6139	0.0038
AMP		
Attention	3.3670	-0.0018
Relevance	4.0978	-0.0249***
Confidence	3.4879	0.0278***
Satisfaction	3.4140	0.0184**

Note. n = 165.

**p<.05. **p<.01. ***p<.001.*

Study 5: Non-Faculty Factors Related to Students' End of Course Ratings of Their Professors

When interpreting student ratings of college instruction one should be aware that ratings can be systematically higher or lower because of factors that are unrelated to instruction (Haladyna & Hess, 1994). If bias is present, but undetected, the validity of the interpretations is jeopardized. For the undergraduate assessment course, nine student variables were identified that could potentially bias student ratings of instruction: major area of study (Major), number of semester hours completed since admission to the College of Education (Training), prior teaching experience (Experience), current number of semester hours (Load), number of hours employed (Work), the length of time the student had wanted to become a teacher (Aspire), the number of years the student planned to teach after graduation (Career), comfort with mathematics (Math) and experience with computers (Computer). The purpose of this study was to explore four questions: (a) Do the mean levels of the ratings of instruction vary across class sections? (b) Do the mean levels of the student variables vary across class sections? (c) Are the student variables related to the student ratings of instruction? (d) Do the relationships among the student variables and the student ratings of instruction vary across class sections?

Sample and Procedures

Undergraduate students (n = 500) enrolled in 23 sections of an undergraduate assessment course participated in the study. In the last week of class, students completed the eight SUSSAI items using one of three response scales. After preliminary analyses, an overall rating was obtained by calculating the mean of the eight items (Cronbach alpha internal consistency reliability index = .98). The decision to include data from all three response scales was based on a series of preliminary analyses showing no statistically significant mean differences among forms on the composite SUSSAI rating (p>.05). Analyses were carried out using two-level hierarchical models with students nested within class sections. Data for the student variables were obtained with biographical questions that accompanied the end of course evaluation instrument. The questions were formatted in an objective style with five response choices with one representing the lower and five representing the higher amount of the attribute (e.g., How many semester hours have you completed since admission to the College of Education? 1 = I have not yet been admitted, 2 = 0-9, 3 = 10-30, 4 = 31-45, and 5 = 46 or more).

Results and Discussion

The composite ratings of instruction varied across class sections by an amount that was statistically significant and practically important. A summary of the distribution of section means on the SUSSAI and the results of the likelihood ratio test are presented in Table 6. The minimum section mean of 2.79 indicates an average response between fair and good, while the maximum section mean of 4.69 indicates an average rating between very good and excellent. The class sections also varied as a function of the student variables. The sections differed by a statistically significant amount on the student variables of Training, Experience, Load, Work, Aspire, and Computer, but not on Math and Career.

The distributions of section means for these variables as well as the results of the likelihood ratio tests are presented in Table 7. As an example, consider the variable Work.

Table 7
Summary of the Distribution of Section Means and Likelihood Ratio Tests for Random Variability Across Class Sections

	M	SD	Minimum	Maximum	χ^2 (1)
SUSSAI	4.09	0.40	2.79	4.69	34.8*
Student variables					
Training	3.69	0.49	2.82	4.80	41.2*
Experience	3.17	0.58	1.84	4.05	41.3*
Load	4.02	0.48	2.81	4.63	85.6*
Work	2.98	0.59	2.09	4.35	26.9*
Aspire	3.78	0.33	3.05	4.37	4.8*
Career	4.13	0.29	3.50	4.65	0.0
Math	2.99	0.34	2.47	3.69	0.0
Computer	3.79	0.31	3.11	4.43	6.6*

Note. Training: number of semester hours completed since admission to the College of Education (1 = I have not yet been admitted, 5 = 46 or more); Experience: prior teaching experience (1 = I have not observed in a classroom, 5 = I have taught unsupervised); Load: current number of semester hours (1 = 3, 5 = 16 or more); Work: number of hours employed (1 = none, 5 = 31 or more); Aspire: length of time the student had wanted to become a teacher (1 = I still am not sure I want to be a teacher, 5 = Since before high school graduation); Career: number of years the student planned to teach after graduation (1 = At this point I do not plan to be a classroom teacher after I graduate from college, 5 = At this point, I plan a career as a teacher); Math: comfort with mathematics (1 = I have difficulty understanding mathematically related concepts and processes, 5 = I am very comfortable); and Computer: experience with computers (1 = I have no prior experience with computers to 5 = I am a regular computer user and am familiar with a variety of computer applications).
**p < .05.*

The smallest section mean was 2.1 where a rating of 2 corresponds to 1-10 hours of weekly employment. The largest section mean was 4.3 where a rating of 4 corresponds to 21-30 hours of weekly employment. Sections also varied in terms of students' major area of study. For example, the percent of elementary majors ranged from 0% to 100%, the percent of secondary majors ranged from 0% to 64%, the percent of special education majors ranged from 0% to 92%, and the percent of performance area majors ranged from 0% to 71%. Since the classes differed on both ratings of instruction and student variables, attention was turned to whether the student variables were related to the ratings.

Table 8 contains a summary of the tests for relationships among the student variables and the ratings of instruction. When the ratings of instruction were modeled using each of the student variables individually, it was found that Work was associated with lower ratings. The estimate, however, was -.06. A one-unit change in Work (e.g., moving from the rating of 21-30 hours per week to the rating of 31 or more hours per week) leads to a predicted rating of instruction that is only .06 points lower. Consequently, the effect is small. When the overall ratings of instruction were modeled using all of the student variables simultaneously, Work was again found to be negatively related to the overall ratings (estimate = -.07). The other student variables were not found to have statistically significant relationships with the ratings when they were entered in the model individually or when they were entered in the model together. Finally, there was no statistically significant evidence of variability in effects across class sections (likelihood ratio χ^2 (9) = 4.1, p > .05).

To provide a descriptive supplement to these analyses a standard multiple regression was run which modeled the composite ratings as a function of the nine student variables. The R^2 for this model was only .051, which underscores the lack of strength in the relationship between the ratings and the set of nine student variables. Although the students variables clearly differed across sections of this course, these variables do not appear to bias the ratings to a notable degree. This conclusion is limited to this particular course and the set of factors examined. It is recommended that further studies be conducted which look at more potential biases and a wider range of courses.

Table 8

Tests of Fixed Effects in Two Level Models of Student Ratings of Instruction

	1 Variable in Model				All Variables in Model			
Variable	Num df	Den df	F	*p*	Num df	Den df	F	*p*
Major[a]	4	55	0.87	.49	4	55	0.66	.62
Training	1	22	0.00	.96	1	22	0.02	.88
Experience	1	22	1.61	.22	1	21	0.96	.34
Load	1	22	0.34	.57	1	22	1.24	.28
Work	1	22	5.29	.03	1	22	5.34	.03
Aspire	1	22	1.56	.22	1	21	1.10	.31
Career	1	22	1.68	.21	1	21	0.07	.80
Math	1	22	0.07	.80	1	20	0.05	.83
Computer	1	22	1.86	.19	1	20	1.60	.22

[a]*See note to Table 6 for variable and response definitions.*

Study 6: Influence of Response Scale Format on Students' Ratings of Professors

The original SUSSAI contained a rating scale that went from 1 = Excellent to 5 = Poor. Shortly after it was developed, the rating scale was changed to 1 = Poor through 5 = Excellent. Does the response scale format on the SUSSAI affect students' ratings of faculty and courses? Nunnally (1978) regards any measurement scale as a convention to be agreed upon by researchers, and he further states that a necessary criterion for adequacy of a response scale format is that it be not simply arbitrary, but defensible. Various aspects of Likert-type scales used with evaluation measures have been studied, e.g., effects of positive and negative wording on scale use (Bergstrom & Lunz, 1998), and respondents' definition and use of labels relative to their position within the scale (Klockars & Yamagishi, 1988). Previous research on effects of the direction and format of the response scale has produced mixed results. Barnette (1999) concluded the presence of a primacy effect was not supported by results in a study of the interaction of positive versus negative wording and order of response category. Weng and Cheng (2000) found response category order had no substantial influence on participants' responses or characteristics of the scale such as the factor structure. In contrast, Chan (1991) did find the presence of a primacy effect in that order of response influenced participants' choices as well as the instrument factor structure. With regard to differing anchor labels, Chang (1996) found means and standard deviations did not differ across two kinds of labels and concluded that different anchoring labels would not meaningfully influence the results of attitude measures for research purposes. In a study examining true/false versus agree/disagree response labels on an instrument related to characteristics of effective schools, Tesh, McKenzie, and Jaeger (1992) found no difference in the reliability of the scales, amount of time for completion, and respondents' written comments about the labels, although the true/false label had a higher non-completion rate than the agree/disagree option.

The purpose of this study was to examine the effect of the response scale format on students' ratings of course instruction and instructors. This experiment was designed to inform the researchers whether one of these formats would be more likely than another to elicit student responses at the high quality end of the scale or at the low quality end of the scale.

Sample and Procedure

Undergraduate students (n = 511) enrolled in 24 sections of an assessment course participated in the study during semesters from fall, 1995 to spring, 1997. The sample, education students from the same state university in Florida, were majoring in the following areas: 43% elementary, 30% secondary, 20% special education, 6% performance area (art, music, etc.) and 1% other.

Forms A, B, and C of the SUSSAI used in this experiment were direct adaptations of the State University System Student Assessment of Instruction (SUSSAI) evaluation form that asks for student responses to very global statements about instructional content and delivery. For all three forms, the item stems remained unchanged, Form A used the original SUSSAI response scale which ranged from 1 Excellent to 5 Poor; Form B used the current scale that ranged from 1 Poor to 5 Excellent; and Form C used a Likert scale that ranged from 1 Strongly Disagree to 5 Strongly Agree. Students within each class section were randomly assigned to one of the three response formats, with 172 students completing Form A, 165 students completing Form B, and 174 students completing Form C. This study examined responses on all forms that were completed by students at week 15. A series of one-factor ANOVAs were used to compare students' responses on the three forms.

Results and Discussion

The means and standard deviations for each of the eight items across the three response-scale format forms are presented in Table 9. A series of 1-Way ANOVAs comparing the responses for all eight items of the SUSSAI across three forms revealed no significant differences, after a Bonferroni adjustment, on the means and variances for the eight items. No significant differences (ps >.05) were observed in student responses associated with the three response-scale formats presented to students on these evaluation items. It was observed that students were more likely to choose the highest quality rating using the "strongly agree" label rather than with the "excellent" label. This observation should be examined further in a study designed to address this tendency directly. Further study might also address why some variables in course evaluation and not others might be influenced by response-scale format.

Table 9
Means and Standard Deviations of Responses Across Scale Formats

Item	Form A[a] '1' Excellent to '5' Poor		Form B[b] '1' Poor to '5' Excellent		Form C[c] '1' Strongly Disagree to '5' Strongly Agree	
	M	SD	M	SD	M	SD
1. Description of course objectives and assignments	4.15	0.96	4.18	0.84	4.22	0.98
2. Communication of ideas and information	3.95	1.06	4.04	0.93	3.98	1.04
3. Expression of expectations for performance in this class	4.06	0.98	4.19	0.86	4.22	0.99
4. Availability to assist students in or out of class	4.17	1.01	4.30	0.92	4.29	0.99
5. Respect and concern for students	4.13	1.07	4.23	1.01	4.25	1.07
6. Stimulation of interest in the course	3.72	1.20	3.95	1.02	3.91	1.14
7. Facilitation of learning.	3.92	1.06	4.10	0.93	4.03	1.06
8. Overall assessment of instructor	4.02	1.08	4.20	0.97	4.12	1.06

[a] *n = 172;* [b] *n = 165;* [c] *n = 174.*

Discussion

Logical content analysis of the items on the State University System Student Assessment of Instruction (1995) instrument suggests, individually, they are similar to dimensions of instruction found in earlier research. The factor analysis studies indicate that, as a collection of statements, the instrument reflects a general construct: students' global reaction to instruction. The factor structure remained stable with both undergraduate students on campus and with web-based distance graduate students, indicating its versatility with differing student groups. In addition, the SUSSAI appears to be stable across item and response formats so that some adaptability to suit the context of courses (e.g., a distance education delivery format) appears possible if accompanied with the necessary validity and reliability estimates.

Results from the studies comparing students' ratings multiple times across a semester, however, suggest that the variable measured by the global items on the SUSSAI may reflect more generalized motivation of students, a relatively stable personality characteristic that helps to explain their general attitudes about school and instruction at the college and university level, rather than their situation motivation that explains changes in their attitudes relative to a particular course and instructor. This finding prompts the policy question, how valid are students' general perceptions of instruction as an indication of the quality of a particular course or instructor? Likewise, how valid is it to publish students' general motivations about instruction in the library as an indication of the quality of a particular course or faculty member? More research with a larger selection of students, courses, colleges, and universities is undoubtedly warranted.

Data from the eight items of the SUSSAI indicate variability in responses from class to class and differences were found from class to class on characteristics not related to instruction. It might be reasonable to expect differences observed in students' ratings of instruction to be associated with non-instructional student characteristics. With the exception of the number of hours employed outside of school, it was observed that the student characteristics examined in this study were poor predictors of students' end of term ratings of instruction. It was concluded that there was no substantial explanation of students' ratings of instructor by the nine variables included in this article. This is consistent with earlier research summarized by Marsh (1987), Feldman (1997), and McKeachie (1997) who have concluded that a variety of variables that are suspected to influence student ratings, have little effect on them. In light of the changing nature of undergraduate students (e.g., working more hours outside of studies, age, family unit), the caution by Haladyna and Hess (1994) that even small effects of a characteristic on students' ratings should be considered in interpreting the data. In the study by Carey et al. (1992), the best predictor of end of course ratings on the College's SET and the AMP instruments was students' initial ratings of the course and instructor on the first day of the class.

Even though research into best practice for faculty evaluation has been ongoing for years, there is renewed interest in both methods and procedures, especially related to formative evaluation and course improvement. Differences in current students' characteristics and needs, rapidly changing instructional technology, diminished resources, and the current focus on value-added assessment are all considerations in current models for assuring accountability in instruction. Perhaps policies related to the newer models of value-added, pre-post measures of students' achievement in a course should be accompanied by policies related to pre-post measures of students' attitudes of course quality. Certainly, research into the viability of new policies, instruments, and procedures is warranted.

Limitations

The samples used in these studies are convenience samples composed of students from the same university selected by the chance of enrollment in a particular education course across several terms. *Measurement for Teachers* is a required course for every student enrolled in an undergraduate teaching program. The participants in the distance sample were also comprised of students enrolled in particular web-based courses in Education and Arts and Sciences. It would have been preferable to draw the sample from multiple courses within a college, multiple colleges, and multiple universities.

Author Note

Portions of this paper were presented in a symposium at the 1997 annual meeting of the Florida Educational Research Association in Orlando, Florida. James White, one of the co-authors, is currently one of the editors of Florida Journal of Educational Research. The other authors are or have been with the University of South Florida, College of Education, Department of Measurement and Research, the same affiliation as a second editor of FJER.

Correspondence concerning this article should be addressed to Tary L. Wallace, College of Education, Department of Educational Research, Technology and Leadership, University of Central Florida, Orlando, FL.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (pp.9-48). Washington, DC: American Educational Research Association.

Arreola, R. A., & Aleamoni, L. M. (1990). Practical decisions in developing and operating a faculty evaluation system. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice.* New Directions for Teaching and Learning, no 43. New York: Jossey-Bass.

Barnette, J. J. (1999, April). *Likert response alternative direction: SA to SD or SD to SA: Does it make a difference?* Paper presented at the annual meeting American Educational Research Association, Montreal, Canada.

Bergstrom, B. A., & Lunz, M. E. (1998, April). *Rating scale analysis: Gauging the impact of positively and negatively worded items.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Beverly Hills, CA: Sage Publications.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1994). *Hierarchical linear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Carey, L. M. (1990, November). *Development and validation of the Academic Motivation Profile.* Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee, FL.

Carey, L. M. (2001). *Measuring and Evaluating School Learning* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Carey, L. M., Carey, J. O., Dedrick, R. F., Wallace, T. L., Kushner, S. N. (1994, November). *Students' evaluations of courses: What do they mean?* Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Carey, L. M., Carey, J. O., & Pearson, L. C. (1992, April). *Validity problems in using end of course ratings to infer instructor and course quality.* Paper presented at the American Educational Research Association conference, San Francisco, California.

Carey, L. M., Dedrick, R. F., Carey, J. O., & Kushner, S. N. (1994). Procedures for designing course evaluation instruments: Masked personality format versus transparent achievement format. *Educational and Psychological Measurement, 54,* 134-145.

Cashin, W. E. (1994). Student ratings of teaching: A summary of the Research. In K. A. Feldman & M. B. Paulsen (Eds.), *Teaching and learning in the college classroom,* ASHE Reader Series. Needham Heights, MA: Simon & Schuster.

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

Centra, J. A. (2000, June). *What contributes to student self-reported course learning*? Paper presented at the American Association of Higher Education Assessment Conference, Charlotte, NC.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement, 51,* 531-540.

Chang, L. (1996, April). *Dependability of anchoring labels of Likert-type scales.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching effectiveness. *Structural Equation Modeling, 7,* 442-460.

Chickering, A.W. & Gamson, Z. F. (Eds.). (1991). *Applying the seven principles for good practice in undergraduate education.* San Francisco: Jossey-Bass.

Crittendon, K. S. & Norr, J. L. (1973). Students' values and teacher evaluation: a problem in person perception, *Sociometry, 36* (2), 143-151.

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

d'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52,* 1198-1208.

Dedrick, R. F., Carey, L. M., Carey, J. O., Wallace, T. L., Greenbaum, P. Ferron, J. M., & Kushner, S. N. (1995, November). *Modeling individual change in multiple dimensions of a course evaluation instrument.* Paper presented at the meeting of the Florida Educational Research Association, St. Petersburg, FL.

Dilts, D. A., Haber, L. J., & Bialik, D. (1994). *Assessing what professors do: An introduction to academic performance appraisal in higher education.* Westport, CT: Greenwood Press.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice.* Bronx, NY: Agathon Press.

Finaly, E. and Neumann, Y. (1985). The measurement and meaning of students' satisfaction with instruction. *Journal of Instructional Psychology, 12*(1), 11-18.

Gagné, R. M. (1985). *The conditions of learning (*4th ed.). New York: Holt, Rinehart and Winston.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182-1186.

Haladyna, T. & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35,* 669-687.

Joreskog, K. G., & Sorbom, D. (1995). *LISREL 8 user's reference guide.* Chicago, IL: Scientific Software.

Kaplan, P. S. (1990). *Educational psychology for tomorrow's teachers.* New York: West Publishing Company.

Keller, J. M. (1987a). Development and use of the ARCS model of instructional design. *Journal of Instructional Development, 10(3),* 2-10.

Keller, J. M. (1987b). *The systematic process of motivational design. Performance and Instruction, 26(*9), 1-8.

Klockars, A. J. & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25(*2), 85-96.

Legg, S. M. & Cunningham, W. (January, 1995). *Pilot test of the university-wide teacher evaluation form.* Gainesville: University of Florida, Faculty Committee on Teacher Evaluation, Office of Instructional Resources.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 257-303.

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 91,* 285-296.

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

McKeachie, W. J. (1997). Student ratings: The validity of use. American Psychologist, 52, 1218-1225.

Nunnaly, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Pintrich, P. R. (1990). Implications of psychological research on student learning and college teaching for teacher education. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 826-857). New York: Macmillan.

Slavin, R. E. (1991). *Educational Psychology* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.

State University System of Florida Board of Regents. (1995). *State University System Student Assessment of Instruction.* 325 West Gaines Street, Tallahassee, Florida 32399-1950.

Tesh, A. S., McKenzie, C. S., & Jaeger, R. M. (1992, February). *Evaluation of the effectiveness of alternate questionnaire item response options*. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill, NC.

Tiberius, R. G. & Billson, J. M. (1991). The social context of teaching and learning. In R. J. Menges & M. D. Svinicki (Eds.), *College teaching from theory to practice*. New directions for teaching and learning, no.45. San Francisco: Jossey-Bass.

Weng, L. & Cheng, C. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement, 60,* 908-924.