Student Response to Faculty Instruction (SRFI): An Empirically Derived Instrument to Measure Student Evaluations of Teaching

Author Info:

Brian D. Beitzel, Department of Educational Psychology, Counseling and Special Education, SUNY Oneonta.

Correspondence concerning this article should be addressed to Brian D. Beitzel, Department of Educational Psychology, Counseling and Special Education, SUNY Oneonta, 20 Denison Hall, Oneonta, NY 13820. E-mail: brian.beitzel@oneonta.edu

Abstract

The Student Response to Faculty Instruction (SRFI) is an instrument designed to measure the student perspective on courses in higher education. The SRFI was derived from decades of empirical studies of student evaluations of teaching. This article describes the development of the SRFI and its psychometric attributes demonstrated in two pilot field tests. Students in all levels of courses (freshman through graduate) and from many different class sizes participated in these two pilot studies. Response rates were high and the ratings students assigned to the SRFI items corresponded with course characteristics in ways consistent with the empirical literature. Through these two pilot studies the SRFI has shown itself to be a reliable and valid instrument for ascertaining the student perspective on instruction in higher education.

Keywords: student evaluations of teaching, reliability, validity

Student Response to Faculty Instruction (SRFI):

An Empirically Derived Instrument to Measure Student Evaluations of Teaching Faculty in higher education need no introduction to student evaluations of teaching (SETs). Privately and publicly, these instruments are often maligned as a special kind of curse on faculty who have contributed their earnest energies in a sincere effort to be the best instructors they can be. When the end of the term arrives, students are invited to voice their opinions about the course and the instructor—and almost invariably, one or more students (particularly those who lacked personal responsibility for their performance in the course) will not see the instructor in the most favorable light. At many institutions, data from these instruments weigh heavily in tenure and promotion decisions (Baldwin & Blattner, 2003; Lindahl & Unger, 2010). The perception among many faculty is that they could be punished (i.e., with a negative personnel decision) for something they did not do.

A portion of this distaste is justified. Data from SETs are regularly reviewed by peers (e.g., on tenure and promotion committees) and administrators, few of whom have any training in the proper interpretation of these data (McKeachie, 2007). Wilbert McKeachie, a well known and highly respected scholar of college teaching, expresses his view of this situation unequivocally: “The major validity problem is in the use of the ratings by personnel committees and administrators” (McKeachie, 1997, p. 1222). Only a handful of dossier reviewers possess the discernment to recognize what the instrument was designed to measure, what its limitations are, and what evidence exists to demonstrate its validity. When confronted with a veritable mountain of data containing numbers to the hundredths decimal place, one could feel compelled by this ostensible precision to arrive at a conclusion (positive or negative) that may be unwarranted, perhaps without even being aware that such definitive-looking numbers do not represent the certainty we wish they would. This perception of accuracy can result in unjustifiable actions, such as ranking faculty for merit considerations on the basis of SET ratings. According to McKeachie, “attempting to compare teachers with one another by using numerical means or medians” is a regrettable use of SET data (1997, p. 1222).

Worse yet, SETs are not always developed by measurement professionals. In these cases, faculty are quite justified for their concerns about the integrity of the instrument. On my campus, a SET was developed and used for 20 years without any formal examination of its validity. Several of the 11 items on this instrument are of suspicious worth—for example, “Classes met regularly as scheduled.” It turns out that those of us who have met every class as scheduled for the entire semester still do not get perfect scores on this item. Thus, because the data for this item are demonstrably inaccurate, the item is of dubious value. (One may be allowed to wonder what other items may be garnering responses that are similarly inaccurate.) Likewise, the item “Methods of assessment were graded fairly” poses a problem for instructors who hold less-than-ambitious students to high standards; many times these students seem to believe that grades which are “fair” are those in the “A” range.
There is one further point to make about instruments that are developed by non-experts. Many SETs ask students to respond to a final item that says something like, “Overall, this is a good instructor.” There is

nothing at all wrong with this kind of "global" item. However, there is an extensive cognitive-science literature demonstrating indisputably that prior-knowledge activation (or "priming") has an effect on subsequent responses (e.g., Bransford & Johnson, 1972; Kinchla, 1992; Logan, 1980; Masson, 1995; McNamara, 1992; Ratcliff & McKoon, 1995; Schacter & Cooper, 1995). In the case of SETs, this means that whatever ideas students have been asked to think about in earlier items could influence how they respond to later items. Specifically, asking students to think about whether classes met as scheduled and whether they received the grade they thought was fair "primes" them for the final global item about the overall quality of the instructor. Regardless of the rigor of the course or the practical value of the required activities (ideas which are not activated by questions on this instrument), students' opinions about the overall quality of teaching are influenced by the ideas most recently activated. After being prompted to remember, among other things, that the teacher held class every day and gave high grades for all student work, the trajectory leads toward marking that person as a good teacher overall. Now it should be easy to see how one could legitimately question the data for even a global item if it follows other items asking about teaching behaviors which even the most ineffective of instructors can easily produce. I have never seen an empirical test of this in the SET literature, but it would be an excellent (perhaps shocking) study.

Because SETs have become necessary in higher education, we must strive to advocate for the highest quality instruments we can achieve within our inherently political environments. SETs that are generated exclusively through group discussion and idea-sharing—without access to empirical data or measurement expertise—are unequivocally inadequate. Higher education, of all places, ought to pursue such endeavors in the most informed manner possible.

Development of the SRFI

The purpose of this article is to describe the development and psychometric properties of the Student Response to Faculty Instruction (SRFI, pronounced *SIR-fee*), a new SET derived from many empirical studies of SETs conducted over the last several decades. Because of the above-described problems with our existing instrument, we (the Committee on Instruction, of which I was chair) elected to abandon the existing instrument and develop a new one.

Purpose of the SRFI

For practical reasons, the SRFI was designed to be an instrument providing summative feedback about a course, from the students' perspective, to both the instructor and administration (chair, dean, etc.). Such information could also be used formatively for future course improvement; however, the principal goal was to obtain high quality summative data.

Guiding Principles

At the outset, our chief goal was to develop an instrument that would be general enough to cover all types of courses at our institution (e.g., freshman, senior, graduate, large, small, general education, distance-learning) to improve the concurrent situation in which some faculty felt they were forced to administer an instrument that was not fully compatible with their course.

A second goal was to have as empirical a rationale as possible for inclusion and exclusion of any potential item to minimize the disagreements that would inevitably result from faculty advancing their own ideas of what should be included or excluded.

To help us meet this second goal, we settled on Feldman's (2007) categorization scheme as our guiding framework. Essentially, he uses two empirical mechanisms to determine which dimensions of teaching are the most important for identifying effective instruction. First, some specific qualities of an effective teacher should correlate highly with student achievement. On average, the courses with the highest performing students are likely taught by instructors with at least some of these superior qualities (Cashin & Downey, 1992). For SETs, this translates to looking for the strongest positive correlations between student achievement (e.g., final exam scores) and specific SET items. Thus, the items that have the strongest positive correlations with student achievement are expected to be the best items for identifying effective teachers. Such items provide us with the best opportunity to escape the priming effect that I described earlier with regard to global items. In Feldman's analysis, items measuring the "Clarity and Understandableness" of the teacher had one of the strongest positive correlations with student achievement. In contrast, items measuring the "Nature and Usefulness of Supplementary Materials and Teaching Aids" had a negative correlation with student achievement.

Feldman's second mechanism for empirically discriminating teacher quality is student responses to global items (e.g., "This is an excellent instructor"). The rationale for this is that if a teacher is an excellent instructor overall, that person must surely also be rated highly on some of the more specific (non-global) SET items. The SET items that meet this second mechanism are those that correlate highly with one or more global items. For example, Feldman found that items addressing the "Teacher's Availability and Helpfulness" correlated most highly with global items. Conversely, items measuring the "Teacher's Stimulation of Interest in the Course and Its Subject Matter" had the weakest correlation with global items.

When selecting potential items for a new instrument, one could rather easily be justified by including an item when it complies with either one of Feldman's mechanisms. However, we specified that *both* mechanisms must be operative in order for us to include an item on our new instrument. That is, in order to be considered for inclusion on our new instrument, a potential item must conceptually fit within one of Feldman's categories having the strongest positive correlations with (a) student achievement *and* (b) global ratings. Because Feldman's mechanisms were founded upon decades of empirical work, we had a solid empirical basis for item selection.

Process

After our selection criteria were established, we generated a pool of items that followed our guiding principles described above. These items were derived from examining other similar instruments, feedback from our campus colleagues via an open-ended survey, Feldman (2007) and other published literature (e.g., Marsh, 1982a; Spooren, Mortelmans, & Denekens, 2007), and our existing campus instrument. The SEEQ (Students' Evaluations of Educational Quality; Marsh, 1982a) is a well known and highly respected instrument for ascertaining student opinion of instruction in college classes. Spooren et al. (2007) developed an instrument that has both a theoretical and empirical basis, but was piloted on a limited set of courses; nonetheless, their sub-dimensions were useful in constructing an initial item pool consistent with Feldman's (2007) framework. We modified the items as necessary for clarity and to avoid copyright infringement.

We did not seek to attain coverage of all of Feldman's categories. Rather, we sought to design an instrument that was in the range of 7-12 items that met the above-mentioned criteria. After assembling an initial item pool, we subjected our fledgling instrument to a pilot test.

Pilot Study 1

Method

Through my committee's informal network of colleagues around campus, we recruited 42 classes of diverse types (large, small, freshman, upper-division, graduate) to participate in the first pilot study. Of the classes that participated, 7 were 100-level, 20 were 200-level, 10 were 300-level, and 5 were graduate classes. The median enrollment for these classes was 20 students. A third party (a secretary) was the only person who maintained and had access to the participant list. All communication between me and the participating faculty was directed through her to maintain confidentiality.

Pilot Instrument 1

The instrument for Pilot Study 1 had 22 items to be rated on a 5-point scale ("Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree"). Many of these items were being tested for inclusion on the final instrument; these are referred to throughout this article as "pilot items"; the subset of pilot items that were eventually included on the final version of the SRFI are referred to in this article as "SRFI items."

Validation Items.

In addition to the pilot items, several other items were included on the pilot instrument for validation purposes, described in the succeeding paragraphs. These items are referred to throughout this article as "validation items."

Student Motivation.

One of the strongest correlates with SET ratings is students' prior interest in the subject (Marsh, 2007). The more motivated a student is in a specific course, the higher that student's grade is likely to be—and the more the student should rate the instructor favorably (Cashin, 1995). To test this assumption, we included three motivational items on our pilot instrument to look for relationships between them and the pilot items.

Quality of Instructional Goals.

Similarly, we included one item to determine whether the instructors who had course goals that were higher than simply memorization of the material would receive different ratings than courses in which students felt their principal task was memorization. In the SET literature, courses that are more difficult tend to receive higher student ratings (with the exception of extremely difficult courses; Marsh, 2007).

Knowledge Gain.

Students who believe they have learned a lot in the course are more inclined to rate a professor favorably than students who do not perceive they have learned much in the course (Cashin, 1995; Feldman, 2007; Stapleton & Murkison, 2001). One item related to perceived knowledge gain was included on the pilot instrument to test for this relationship.

Global Items.

We included three global validation items (beginning with the word "Overall") to assess the relationship between these global items and the pilot items (Feldman's second mechanism for determining teacher quality from student ratings, discussed earlier). These global items were presented as the last items on the pilot instrument.

Results and Discussion

Response Rate.

Overall, 775 students (81%) responded to the pilot instrument. By practically any standard, this is a very satisfactory response rate. Among the 37 undergraduate classes, 738 students (80%) responded, and from the 5 graduate classes 37 students (95%) responded.

Item Reduction.

One goal of Pilot Study 1 was to narrow down the pool of items to a final size of 7-12 items (see the appendix for the final set of items). To determine which items to keep, a stepwise regression analysis was conducted, using a criterion of p = 0.05 for entry and p = 0.10 for exit. The dependent variable was each of three global validation items. Thus, three regression analyses were conducted. In order to be included on the final version of the SRFI, we determined that a pilot item must have been a predictor in *at least two* of the three regression analyses in *both* Pilot Study 1 *and* (eventually) Pilot Study 2. These stringent criteria were designed to help us select only the highest quality, most reliable items for the final instrument. Multicollinearity is a concern for analyses of this type, due to the expected correlation among typical SET items. To examine this assumption, the coefficients, their standard errors, and the significance tests were visually inspected at each step of these stepwise regression models. Across the entire series of regressions reported in this article (in both Pilot Study 1 and Pilot Study 2), the coefficients and their standard errors were remarkably stable, and there were no cases in which the significance tests were inconsistent across each progressive step of the regression (cf. Kidwell & Brown, 1982). Thus, we can state with confidence that the independence assumption associated with the multiple-regression analyses has been met.

Regression Results.

All regression models had omnibus ps less than 0.0001. The final (stepwise) models contained between 5 and 8 predictor variables, demonstrating that we do not need a very large number of items to make a satisfactory prediction about students' views regarding their instructors' overall teaching skill. The final models included at most one item that was not eventually included on the final version of the SRFI; the non-included item was not the same item in all three models. Thus, we did not eliminate any item that the regression analyses uniformly indicated as highly useful in predicting responses to the global validation items. The R^2 values for these final stepwise models ranged from 0.707 to 0.841; therefore, we can account for a modestly strong proportion of the variance in the global validation items with a small subset of the pilot items. The pilot items that entered the regression models but were not included on the final version of the SRFI entered no earlier than Step 4 of the regressions. Thus, we did not eliminate any items that were most predictive of overall teaching quality.

Pilot Items Eliminated. Several pilot items were eliminated on the basis of the regression results because they were not predictors in more than one of the three regression models. For example, the three items, “The instructor was prepared for class,” “The course was well organized,” and “The instructor responded satisfactorily to questions both in and out of class” were predictors for exactly *none* of the three global items and were therefore eliminated from further consideration.

Global-Like SRFI Item. The single global-like item retained on the final version of the SRFI was, “I would recommend this instructor to other students.” This item was the first to enter in each of the three stepwise regressions, with *F*-ratios ranging from 670.09 to 2697.83, and R^2 values ranging from 0.60 to 0.78. Therefore, this item alone captures most of the variance in the global validation items. In fact, one would not be poorly served by asking this question alone!

Reliability.

Reliability was calculated in three different ways. The well known, widely used, and often-criticized (cf. Cortina, 1993; Green & Yang, 2009; Sijtsma, 2009) coefficient alpha (Cronbach, 1951) yielded an internal consistency measure of 0.93.

Following the advice of Revelle and Zinbarg (2009), the ω_t statistic was also calculated using the `psych` package from the open-source software R (R Development Core Team, 2011). The result was also 0.93.

Marsh (1982b) recommended that an intraclass correlation is the most appropriate statistic for estimating the reliability of SETs because it anticipates high levels of agreement within each course and different ratings among courses. For Pilot Study 1, the intraclass correlation was 0.93.

Thus, from three different approaches to measuring reliability, the results are all 0.93. For the pilot version of any instrument, this level of reliability is a measurement specialist’s dream.

Uni-Dimensionality.

There has been some interest within the SET literature regarding the dimensionality of SET instruments (Abrami, d’Apollonia, & Rosenfield, 2007; d’Apollonia & Abrami, 1997). To test the dimensionality of the SRFI, the Pilot Study 1 data were randomly split into halves, and an exploratory factor analysis was conducted, using only the SRFI items (after the other pilot items had been eliminated) and extracting only factors with eigenvalues greater than 1. Using principal components extraction, only one component was derived, with loadings all greater than 0.72. These high loadings indicate a strong correlation of the items with the extracted component. This component accounts for 67% of the variance among the items. The other half of the data were then subjected to a confirmatory factor analysis, constraining the extraction to a single component, with similar results (loadings of .80 or greater and 72% of the variance explained).

With the eliminated pilot items included, the single derived component accounts for less of the variance among the items (63%) than the above models. Thus, the reduced size of the SRFI is warranted by this analysis.

Validity.

As stated earlier, the SRFI was developed using constructs having an empirically verified relationship with high quality instruction. This process gave us our best chance at having *construct validity*.

As a proxy for *concurrent validity,* we examined the relationship between our validation items and the SRFI items. The validation items were modeled after the constructs described in Feldman (2007) and concerns described throughout the SET literature. The phrasing of the validation items very closely matched corresponding items in the Instructional Development and

Effectiveness Assessment (IDEA; Hoyt & Cashin, 1977). We also tested against such variables as class size and average grade.

Relationship with Instructional Variables. Some validation items represent variables that are at least somewhat under the instructor's influence. These include the extent to which students are challenged to do more than memorize material and the amount of knowledge students perceive they have gained within a course.

Quality of Instructional Goals.

A regression analysis with the quality of instructional goals (from the item "The instructor required students to do more than just memorize course material") as the dependent variable resulted in a model with 7 predictors and an R^2 of 0.568, $F(7, 755) = 141.752$, $p < .001$. All beta coefficients for the three SRFI items included in the model were positive, showing that SRFI ratings tend to be somewhat higher when the instructor requires students to do more than just memorize course material.

Knowledge Gain.

To analyze the relationship of SRFI items with student-perceived knowledge gain, a regression analysis was conducted with the item "I learned a lot in this course" as the dependent variable. This stepwise regression resulted in a model with 8 predictors and an R^2 of 0.611, $F(8, 753) = 148.096$, $p < .001$. All beta coefficients for the six SRFI items included in the model were positive, showing that SRFI ratings tend to be higher when students believe they have learned a lot in the course.

Relationship with Non-instructional Variables.

Some validation items represent variables that lie almost entirely outside of the instructor's control. If there is no relationship between responses to these factors and responses to SRFI items, it would be practically impossible to argue that these constructs are influencing SRFI ratings.

Student Motivation.
Our three validation items related to student motivation asked students about their (a) prior interest in the subject matter, (b) desire to enroll in a course with this instructor, and (c) experience in the course being pleasurable. For each of these items, students responded with the full range of possible values (0–4) across all courses in the aggregate; the aggregate median for these items, respectively, was (a) 3, (b) 2, and (c) 3. In a series of regression analyses with each of these validation items as the dependent variable, three SRFI items did not even enter the models. The other four SRFI items entered the models as predictors, sometimes at the first step, but never at the first step for all three models.
Not surprisingly, the conclusion is that we cannot rule out student motivation as a factor in SRFI ratings. However, we can also say that motivation does not appear to be the principal influence across the board.

Relationship with Other Variables.

There is some inconsistency in the SET literature as to whether class size and course grade make any difference in SET responses. To check this relationship with the SRFI, our third-party coordinator (to ensure anonymity) provided me with course-enrollment and student-performance data for each participating course.

Some articles in the SET literature report student-level correlations, and others report class-level correlations. Since there is no "correct" unit of analysis for SET investigations (Feldman, 2007; Wigington, Tollefson, & Rodriguez, 1989), both are provided (where appropriate) in Table 1. This is instructive, because the increased variability (and increased *n*) in the student-level correlations yields more statistically reliable results; in fact, only two of the 21 class-level correlations are statistically different from zero!

Class Size.

The typical finding in the SET literature—for those studies where any non-zero relationships are found—is that larger classes yield lower SET ratings (Feldman, 1984; Marsh, 1987). For the SRFI, this is somewhat true, although the correlations are quite weak to nonexistent. The only exception to this is SRFI item #6, which had a class-level correlation of -.49 with enrollment. As can be seen from the appendix, item #6 has to do with the instructor caring about the progress of each student in the course; therefore, a negative correlation is expected here on practical grounds alone.

Average Grade.

The relationship between student achievement and SET items is well established in the literature to be a correlation of about 0.30 or less (Feldman, 2007; Marsh & Roche, 1997). To look for relationships between SRFI data and student achievement (Feldman's first mechanism), the class average, as a percent, was correlated with each SRFI item. The resulting correlations with SRFI items are given in Table 1. As these correlations show, there is no significant relationship between SRFI ratings and the average grade for the course. Most importantly, there is no basis in these data for claiming that "easy" courses get higher ratings.

In this sample, the correlation between class size and average grade was -.46, $p < .01$. Thus, the larger classes tended to have lower average grades, which is not a surprising outcome given that the three largest classes in this sample were 100-level classes.

Course Level.

There is some indication in the SET literature that students in advanced courses provide ratings that are more favorable (Aleamoni, 1999; Cashin, 1995). The SRFI correlations related to this are in Table 1. There is at best a weak tendency for students in more advanced courses to rate their instructors more favorably. This is particularly good news for instructors of freshman classes, who often fear lower ratings from unengaged first-year students.

Pilot Study 2

Before recommending the SRFI for campus-wide adoption, it was subjected to a second pilot test. The purpose was to show that the results from Pilot Study 1 could be replicated.

Method

The procedure for Pilot Study 2 was identical to that of the first pilot study. We obtained results from 33 classes: 12 100-level, 12 200-level, 7 300-level, and 2 graduate classes. The median enrollment for these classes was 24 students.

Pilot Instrument 2

The instrument for the second pilot study was identical in form to Pilot Instrument 1. Pilot Instrument 2 retained the validation items from Pilot Instrument 1 but eliminated the items that the regression analyses had not shown to be useful in Pilot Study 1.

Results and Discussion

Response Rate.

In this study, 638 students (87%) responded to the pilot instrument. This is a highly satisfactory response rate. From the 31 undergraduate classes, 616 students (86%) responded, and 22 students (96%) from the 2 graduate classes responded.

Regression Results.

Three stepwise regressions were again conducted, using the same dependent variables and criteria as in Pilot Study 1. Again, all regression models had omnibus *p*s less than 0.0001. The final models included between 5 and 6 predictor variables, all of which were SRFI items. The R^2 values for these final stepwise models ranged from 0.639 to 0.816; as in Pilot Study 1, we can account for a modestly strong proportion of the variance in the global validation items with the SRFI items.

Items Eliminated. There were no pilot items eliminated on the basis of the regression analyses. All of the pilot items retained from Pilot Study 1 functioned as expected.

Reliability.

Coefficient alpha (Cronbach, 1951) for the SRFI items was 0.92, practically identical to the finding from those same items in Pilot Study 1. The second reliability estimate calculated, ω_t, was 0.92, again identical to coefficient alpha and almost identical to the same metric in Pilot Study 1. The intraclass correlation for Pilot Study 2 was also 0.92. Thus, with virtually the same strong results across both of these pilot studies, the SRFI has shown itself to have impeccable reliability in these two samples.

Uni-Dimensionality.

A confirmatory factor analysis was conducted, constraining the extraction to a single component (due to the clear results of a single component from the factor analysis in Pilot Study 1). Once again only one component was extracted, with loadings of 0.748 and greater. These high loadings indicate a strong correlation of the items with the derived component. This component accounts for 68% of the variance among the items.

Validity.

The validity tests from Pilot Study 1 were repeated. These tests examined the relationship between SRFI responses and various course and student characteristics.

Quality of Instructional Goals.

The stepwise model for quality of instructional goals resulted in a model with 4 predictors and an R^2 of 0.349, $F(4, 633) = 84.997$, $p < .001$. All beta coefficients included in the model were again positive, showing that SRFI ratings tend to be higher when the instructor requires students to do more than just memorize course material.

Knowledge Gain.

The stepwise regression for student-perceived knowledge gain resulted in a model with 4 predictors and an R^2 of 0.520, $F(4, 633) = 171.322$, $p < .001$. All beta coefficients were again positive, showing that SRFI ratings tend to be higher when students believe they have learned a lot in the course.

Student Motivation.

The three stepwise regressions for student motivation were again conducted with the same three validation items as dependent variables: (a) prior interest in the subject matter; (b) desire to enroll in a course with this instructor; and (c) experience in the course being pleasurable. For each of these items, students responded with the full range of possible values (0–4) across all courses in the aggregate; the aggregate median for these items, respectively, was (a) 3, (b) 2, and (c) 3. The final stepwise models for these three regressions had the following characteristics, respectively: (a) 3 predictors and an R^2 of 0.08; (b) 3 predictors and an R^2 of 0.40; and (c) 4 predictors and an R^2 of 0.75.

Not surprisingly, we again see that student motivation is related to ratings. The only student-motivation item which the instructor has no control over is students' prior interest in the topic—and that accounts for only 8% of the variance in these data.

Class Size.

As shown in Table 2, the correlations with class size are very weak; most of them are not statistically distinguishable from zero. Therefore, class size does not appear to be influencing SRFI ratings within this sample.

Average Grade.

Students' average grade does not seem to be a heavy influence on SRFI ratings. Most of these relationships are weak, but two correlations bear further inspection. SRFI items #2 and #4 have a modest correlation with average course grade. However, neither of these items point to inflated ratings because of lenient grading. Item #2 is about the instructor demonstrating an interest in the course material; when this happens, it should come as no surprise that students in the class are achieving more. Item #4 has to do with the instructor effectively conveying why the subject is meaningful; would students not be expected to accomplish more when they have been told why the material is important?

In this sample, the correlation between class size and average grade was -.13, $p = .48$. The largest class size in this sample was 60 students, which could account for the nonsignificant finding compared to Pilot Study 1 in which the largest enrollment was 92 and there was a significant relationship between these variables.

Course Level. The correlations with course level shown in Table 2 are mostly weak or nonsignificant. The single exception is item #1, which has to do with the instructor explaining grading policies. The negative correlation indicates that instructors of higher-level courses did not accomplish this as well as their peers teaching lower-level courses. The higher-level courses in this sample likely incorporated more subjective forms of evaluation that were not transparent to students. However, there were only two graduate courses in this sample (in contrast to the five graduate courses in Pilot Study 1), so this correlation of -0.42 could be interpreted as an artifact of the sample when compared to the corresponding 0.08 correlation in Pilot Study 1.

General Discussion

Reliability

Three forms of reliability estimates were used: coefficient alpha, ω_t, and intraclass correlation. All three estimates were uniformly high across both pilot studies, demonstrating that—at least for these two samples—the SRFI is a highly dependable instrument.

One must be cautious, of course, in drawing conclusions beyond the sample data. Reliability is a characteristic of the data set, not of the instrument itself. However, the similarity of all reliability coefficients across both pilot studies (along with the general findings of high reliability within the SET literature) does inspire confidence that reliability will not be a concern moving forward.

Validity

Validity remains the most critically important characteristic of any SET. Yet, as everyone acknowledges, establishing the validity of a SET is extraordinarily challenging. The difficulty stems from the fact that effective teaching is such a complex phenomenon that there exists no uniformly agreed upon, validated instrument for measuring it (Cashin, Downey, & Sixbury, 1994).

In the SET literature, the articles that describe the development of specific SETs typically depend upon expert opinion as the primary source of validation (e.g., Alok, 2011; Barnes et al., 2008; Kember & Leung, 2008). Our approach with the SRFI was to rely upon a broad empirical base of evidence rather than the opinions of a group of experts. This is not to be read as a slight against experts; neither is it intended to cast doubt upon their expertise. Rather, we understood that one of the chief difficulties of utilizing expert teachers for SET development is that they invariably disagree with one another. Additionally, it is not unthinkable that experts might change their minds over time. Therefore, using a body of empirical data that has been validated against outcomes important to effective teaching—like student performance on a final exam—over an extended period of time offers us an attractive alternative path to validity that is unique within the SET literature. One would certainly have to mount a considerable defense to convincingly demonstrate that all the studies Feldman (2007) analyzed were wrong! This approach to validation paid off. By establishing stringent criteria for inclusion of potential items, we gave ourselves the best chance at having a solid instrument on the first attempt. Indeed, the data from the SRFI items in both pilot tests match the pattern of results from Feldman (2007) perfectly. The evidence is unequivocal: Despite generational differences and technological advances, the students of today respond to instruction much like the students of yesterday did.

Interpretation

To help address McKeachie’s (1997) concerns of invalid interpretation of SET data, my committee provided a list of guidelines for interpretation of SRFI data to the campus community. This list was especially intended for use by personnel committees when evaluating SRFI results for purposes of tenure and promotion.

Conclusion

After a careful, empirically driven process of instrument development, the SRFI performed admirably throughout two pilot tests. The reliability and validity (to the extent that validity could be measured) of this instrument in these two pilot studies were highly satisfactory. Already the SRFI has more evidence of reliability and validity than many instruments currently in use. Because the SRFI was piloted in many different disciplines, it can now be recommended for implementation on a broad scale. The relationship between SRFI ratings and course characteristics such as class size and level align with the expectations from the empirical literature. Consequently, the SRFI is a strong competitor to existing instruments—and should be preferred over instruments for which validity has never been formally examined.

Future research on the SRFI should investigate how it functions on a campus-wide scale. Both faculty and administration on my own campus have approved it for campus-wide use as a replacement of our existing (non-validated) instrument; unfortunately, the faculty union leadership blocked its implementation. Therefore, no campus-wide data can be collected at my institution at present. Nonetheless, the SRFI remains a stronger instrument than many of its competitors.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-456). Netherlands: Springer.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166.

Alok, K. (2011). Student evaluation of teaching: An instrument and a development process. *International Journal of Teaching and Learning in Higher Education, 23*, 226-235.

Baldwin, T., & Blattner, N. (2003). Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching, 51*, 27-32.

Barnes, D. C., Engelland, B. T., Matherine, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., . . . Williams, Z. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal, 42*, 199-213.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84*, 563-572.

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology, 86*, 649-657.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

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

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*, 45-116.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-143). Netherlands: Springer.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121-135.

Hoyt, D. P., & Cashin, W. E. (1977). *Development of the IDEA system* (IDEA Technical Report No. 1). Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University.

Kember, D., & Leung, D. Y. P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education, 33*, 341-353.

Kidwell, J. S., & Brown, L. H. (1982). Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and Family, 44*, 287-299.

Kinchla, R. A. (1992). Attention. *Annual Review of Psychology, 43*, 711-742.

Lindahl, M. W., & Unger, M. L. (2010). Cruelty in student teaching evaluations. *College Teaching, 58*, 71-76.

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology, 12*, 523-553.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77-95.

Marsh, H. W. (1982b). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*(1), 47-59.

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

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

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187-1197.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 3-23.

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

McKeachie, W. J. (2007). Good teaching makes a difference—and we know what it is. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 457-474). Netherlands: Springer.

McNamara, T. P. (1992). Theories of priming: I. Associative distance and lag. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1173-1190.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratcliff, R., & McKoon, G. (1995). Bias in the priming of object decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2*, 754-767.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Schacter, D. L., & Cooper, L. A. (1995). Bias in the priming of object decisions: Logic, assumption, and data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 768-776.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

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

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

Wigington, H., Tollefson, N., & Rodriguez, E. (1989). Students' ratings of instructors revisited: Interactions among class and instructor variables. *Research in Higher Education, 30*, 331-344.

Table 1

Pearson Correlations Between SRFI Items and Class Characteristics in Pilot Study 1

Item[a]	Class Size[b]		Avg. Grade[c]	Course Level[d]
	Student[e]	Class[f]	Class	Class
Item 1	0.00	-0.02	-0.12	0.08
Item 2	-0.29**	-0.27	-0.04	0.08
Item 3	-0.03	-0.08	0.04	0.07
Item 4	-0.27**	-0.33*	0.11	0.19
Item 5	-0.16**	-0.21	0.14	0.16
Item 6	-0.36**	-0.49**	0.11	0.30
Item 7	-0.14**	-0.23	0.17	0.19

Note. No student-level data are available for course grade or class year; therefore, only class-level correlations are shown for these variables.

[a] See the appendix for full item text. [b] Range: 2 to 92 students enrolled [c] Range: 71% to 96%

[d] Course Level was coded as follows: 1 = 100-level course; 2 = 200-level course; 3 = 300-level course; 4 = graduate course (our institution does not have 400-level courses). [e] Individual-level correlations between SRFI items and class size. [f] Class-level correlations between SRFI items and class size.

$^{*}p < .05$. $^{**}p < .01$.

Table 2

Pearson Correlations Between SRFI Items and Class Characteristics in Pilot Study 2

Item[a]	Class Size[b]		Avg. Grade[c]	Course Level[d]
	Student[e]	Class[f]	Class	Class
Item 1	0.12**	0.31	-0.04	-0.42*
Item 2	0.02	-0.03	0.51**	0.26
Item 3	0.04	0.18	0.23	-0.17
Item 4	-0.07	-0.16	0.47*	0.31
Item 5	0.03	0.11	0.15	-0.18
Item 6	-0.03	0.05	0.36	0.08
Item 7	0.09*	0.19	0.18	-0.26

Note. No student-level data are available for course grade or class year; therefore, only class-level correlations are shown for these variables.

[a]See the appendix for full item text. [b]Range: 2 to 60 students enrolled [c]Range: 75% to 94%

[d]Course Level was coded as follows: 1 = 100-level course; 2 = 200-level course; 3 = 300-level course; 4 = graduate course (our institution does not have 400-level courses). [e]Individual-level correlations between SRFI items and class size. [f]Class-level correlations between SRFI items and class size.

*$p < .05$. **$p < .01$.

Appendix

Student Response to Faculty Instruction

1. The instructor clearly explained his/her grading criteria, including how final grades in this course will be determined.

2. The instructor was clearly interested in the course material.

3. The instructor presented and explained ideas effectively.

4. The instructor communicated the significance of the subject.

5. Throughout the course, the instructor made it clear what I should learn and accomplish.

6. The instructor was clearly interested in the learning of each student.

7. I would recommend this instructor to other students.