


# A Criterion-Referenced Approach to Student Ratings of Instruction

J. Patrick Meyer<sup>1</sup>  · Justin B. Doromal<sup>1</sup> · Xiaoxin Wei<sup>1</sup> · Shi Zhu<sup>1</sup>

Received: 9 February 2016 / Published online: 13 September 2016  
© Springer Science+Business Media New York 2016

**Abstract** We developed a criterion-referenced student rating of instruction (SRI) to facilitate formative assessment of teaching. It involves four dimensions of teaching quality that are grounded in current instructional design principles: Organization and structure, Assessment and feedback, Personal interactions, and Academic rigor. Using item response theory and Wright mapping methods, we describe teaching characteristics at various points along the latent continuum for each scale. These maps enable criterion-referenced score interpretation by making an explicit connection between test performance and the theoretical framework. We explain the way our Wright maps can be used to enhance an instructor's ability to interpret scores and identify ways to refine teaching. Although our work is aimed at improving score interpretation, a criterion-referenced test is not immune to factors that may bias test scores. The literature on SRIs is filled with research on factors unrelated to teaching that may bias scores. Therefore, we also used multilevel models to evaluate the extent to which student and course characteristic may affect scores and compromise score interpretation. Results indicated that student anger and the interaction between student gender and instructor gender are significant effects that account for a small amount of variance in SRI scores. All things considered, our criterion-referenced approach to SRIs is a viable way to describe teaching quality and help instructors refine pedagogy and facilitate course development.

**Keywords** Criterion-referenced assessment · Teaching quality · Student ratings of instruction · Item response theory

---

✉ J. Patrick Meyer  
jpm4qs@virginia.edu

<sup>1</sup> Curry School of Education, University of Virginia, P.O. Box 400265, 405 Emmet Street South, Charlottesville, VA 22904, USA

## Introduction

Student ratings of instruction (SRI) are an efficient and economical way to obtain feedback on teaching quality. Colleges and universities can quickly collect ratings from thousands of students and provide feedback to instructors thanks to online tools and data collection systems. However, this benefit is often contrasted with contentious debate on the validity and proper use of SRIs. Some argue that SRIs are heavily influenced by factors unrelated to teaching quality (see Johnson 2003), while others maintain that SRIs are adequate measures of teaching quality (see Benton and Cashin 2014). An incomplete review of the vast literature on SRIs could easily support any predilection about their quality and usefulness. A balanced review shows that SRIs are quality measures when they involve multi-item scales to measure multiple dimensions of teaching quality, but also that they should be combined with other measures to provide a more complete picture of teaching quality (Benton and Cashin 2014).

SRIs are typically norm-referenced measures such that an instructor's score takes on meaning by comparing it to the scores of other instructors. It is an approach to score interpretation that is useful for rank-ordering instructors and rewarding the best or rejecting the worst, but it is not a useful way to help instructors become better teachers. If the goal is to improve the quality of instruction, then SRI scores must be criterion-referenced and derive meaning from a teaching framework. Connecting SRI scores to the underlying teaching framework yields specific information about teaching quality and leads to recommendations for refining pedagogy and improving course quality.

Criterion-referenced SRIs may be formative or summative. A formative criterion-referenced SRI is designed to guide faculty development and course improvements. It is the type of SRI that we emphasize in this paper. Summative criterion-referenced SRIs provide data for making high-stakes decisions about promotion and tenure. It involves a formal standard setting procedure (Cizek and Bunch 2007) to define cut-scores that demark different levels of teaching quality (e.g., developing, proficient, advanced). In a bookmark standard setting procedure, a panelist reads through a Wright map (described below) to identify the point where one type of instructor (e.g., developing) is unlikely to receive a rating by students. This point becomes the panelist's cut-score between two levels (e.g., developing and proficient). Standard setting involves a large group of panelists such as a group of experienced instructors. Their judgments are combined in some fashion to establish the cut-scores. Regardless of whether a criterion-referenced test is formative or summative, the underlying content domain or theoretical framework is the criterion, not the group of examinees taking the test.

In this paper, we describe our framework for teaching and the development of an SRI that is aligned to this framework, the Teaching Quality Student Rating Inventory<sup>1</sup> (TQSRI). We then review the literature on SRIs and make note of some persistent threats to the validity of SRIs. Next, we explain our use of item response theory and Wright mapping methods to analyze the data and make a direct connection between SRI scores and the underlying teaching framework. The teaching framework and Wright maps are key aspects of our criterion-referenced approach to score interpretation, but they do not prevent unwanted bias. Therefore, we also studied threats to the validity of our SRI scores to determine whether or not score interpretation is adversely influenced by factors unrelated to teaching quality.

---

<sup>1</sup> The complete measure is available upon request. For brevity, we did not include it in this paper.

## Teaching Framework

Colleges and universities provide high-quality education through high-quality teaching. However, the meaning of “high-quality teaching” depends on instructional design principles adopted by faculty and how well one implements those principles to advance student learning. Thus, quality teaching requires a teaching framework for instructional design, and measurement of an instructor’s ability to implement the teaching framework. Ory and Ryan (2001) explain the importance of these two components by noting that a theoretical basis for teaching quality is necessary for the proper interpretation of SRI scores. They go on to lament the reality that most SRIs have no theoretical basis. SRI questionnaires are often developed in an exploratory manner such that researchers create a large pool of items, collect data, and analyze the result to identify the main components of teaching quality (Marsh 1987; Ory and Ryan 2001). It is much less common for a SRI questionnaire to be developed in a confirmatory approach where the construct of teaching quality is defined prior to the development of the SRI. Although both methods are viable, Benson (1998) writes that exploratory methods mark a weak program of construct validity, whereas a confirmatory approach and testing of rival hypotheses are essential characteristics of a strong program of construct validity.

The construction of our measure was guided by the confirmatory approach. Our dimensions of teaching quality were originally defined by the work of Fink (2013) and Arreola (2007), and advice from expert teachers at the university. We refined our teaching quality dimensions and the items written to measure them over a series of three pilot studies, the third of which is the focus of this paper. In the sections that follow, we define the domains of our teaching quality construct as represented in the most current version of our measure.

### *Organization and Structure*

A central feature of instructional design is writing learning objectives to clearly tell students what they will gain from the course (Brophy 1999; Fink 2013; McKeachie and Svinicki 2006; Wiggins and McTighe 2011). Their use is supported by research on expert-novice differences in the way knowledge is structured in memory. Experts organize information in meaningful ways, which allows them to easily access the information and assimilate new knowledge, but novices tend to treat information as discrete bits of knowledge that must be memorized independently (Bransford et al. 1999). Although expertise takes years to develop, learning objectives expose students to the structure of course content and the cognition it entails. A more direct link between learning objectives and student achievement also exists. Hattie’s (2009) review of meta-analytic studies on student learning found that goals had an effect size of  $d = 0.56$ . A value that was not only large in the absolute sense, but also well above the mean of other factors related to student learning considered by Hattie.

Good learning objectives go beyond statement of central facts to be learned. They also communicate the cognition and thinking involved in the course, the skills students will acquire, and the understandings they will obtain. However, even with quality learning objectives, an instructor’s presentation of course content and instructional activities should not hinder learning. Good instruction models the organization of content by presenting it clearly and without irrelevant digressions or obfuscating discussions (Brophy 1999). Hattie

(2009) found that teacher clarity, an aspect of organization and structure, had a standardized effect size of  $d = 0.75$ .

We define *Organization and structure* as the extent to which the course is organized around learning objectives that are clearly communicated to students, and the extent to which an instructor remains true to the learning objective for the duration of the course. It also involves the clarity of the instructor's presentations, grading procedures, and assignment due dates. Course organization and structure does not involve the extent to which the instructor connected the content to the students' prior knowledge and other coursework. This latter component is included in Academic Rigor because it represents a higher level of organization that is more difficult to implement.

### *Assessment and Feedback*

Assessment and feedback is featured in many approaches to instructional design (see Brophy 1999; Fink 2013; McKeachie and Svinicki 2006). Wiggins and McTighe (2011) emphasize its importance in their backwards design approach by placing it before the development of learning activities and instructional methods. They argue that instructors are better able to plan instruction when they clearly understand the outcomes and evidence of learning. This alignment of learning objectives, assessment, and teaching facilitates student learning (Bransford et al. 1999). Moreover, frequent formative assessment with timely feedback enhances recall of information and promotes self-regulated learning (Pyc et al. 2014). Hattie (2009) found that frequent testing alone has an effect on learning ( $d = 0.34$ ), but this effect is much larger when it is accompanied by feedback, especially feedback aimed at correcting misunderstanding ( $d = 0.73$ ).

Our *Assessment and feedback* dimension involves the alignment of assessment and learning objectives, the frequency of measurement, and the timeliness of information reported to students about their progress. Assessments should measure the content and levels of cognition conveyed in the learning objectives. They may be ungraded assessments that guide instructional decisions and encourage student self-evaluation of learning (i.e., formative assessments). They may also be graded exams or products that require students to demonstrate mastery of the learning objectives (i.e., summative assessments). Assessments should occur frequently enough to monitor student attainment of learning objectives, and feedback should be provided to students quickly enough for them to reflect on their learning and identify objectives in need of more study.

### *Personal Interactions*

Personal interactions contribute to effective teaching and a community-centered classroom (Bransford et al. 1999) through relationships characterized by attachment theory and self-determination theory (see Hamre et al. 2013; Pianta and Hamre 2009). These relationships enhance learning. Brophy (1999, p. 8) writes "students learn best within cohesive and caring learning communities." This claim is supported by his research and the work of others on teacher-student relationships where the effect size is typically large ( $d = 0.72$ ; Hattie 2009).

We define *Personal interactions* as teacher-student interactions that enhance the learning environment. It includes the quality and timeliness of communication as well as the sincerity and concern for student learning demonstrated by the instructor.

## *Academic Rigor*

Prior knowledge and beliefs influence critical thinking (Willingham 2007). Therefore, academic rigor entails a learner-centered teaching environment that pays “careful attention to the knowledge skills, attitudes, and beliefs learners bring to the educational setting” (Bransford et al. 1999, p. 121). Attention to students’ background and prior knowledge helps establish the foundational knowledge that is essential to critical thinking, but teaching itself must go beyond knowledge-centered lectures and involve activities or “thoughtful discourse” (Brophy 1999, p. 19) that engage students in evidenced-based reasoning and problem-solving. For example, once students are properly motivated to learn (Bensley 2010), some meta-cognitive strategies (Willingham 2007) and instruction aimed at identifying students’ misconceptions and requiring them to explain and justify their reasoning help promote critical thinking (Bransford et al. 1999). Motivation ( $d = 0.48$ ), meta-cognitive strategies ( $d = 0.69$ ) and prior achievement ( $d = 0.67$ ) have a strong relationship with learning (Hattie 2009).

Given the complexity of academic rigor, instructional technology provides additional tools for developing deeper learning and understanding (Mayer 2008). For example, videos can bring real-world problems to the classroom and intelligent tutoring systems can provide additional opportunities for feedback (Bransford et al. 1999). However, Clark argues in a seminal article that technology alone will not produce learning (Clark 1983). It must be used in a very intentional way to engage students in higher levels of cognition and critical thinking. For example, Clark (2009) and Kennedy et al. (2014) use Mayer’s (2009) cognitive theory of multimedia learning and accompanying evidence-based instructional design principles to create instructional vignettes that thoroughly consider the cognitive needs of learners.

Our definition of *Academic rigor* concerns learner-centered teaching practices that enrich the learning environment and promote long-term learning and critical thinking. It is culturally-responsive and leverages technology to enhance learning. Although many definitions of critical thinking exist, we define it as reasoning about a problem or issue from multiple perspectives and taking all evidence into consideration when solving a problem or making a decision.

## **Extant Research on the Validity of SRIs**

The literature on SRIs is extensive and complicated by the lack of a universal definition of teaching quality (Clayson 2009; Cohen 1981) and no agreed upon way of measuring it. Marsh (Marsh 1987; Marsh and Roche 1997, 2000) argues that SRIs should be multidimensional and use multiple items to measure each dimension. His view on the dimensionality of teaching quality is consistent with modern instructional design (see Brophy 1999; Fink 2013; McKeachie and Svinicki 2006; Wiggins and McTighe 2011). It is also characteristic of SRIs in K-12 education (see Ferguson 2012) and alternative measures of teaching quality such as standardized observations (see Pianta and Hamre 2009). Measures that involve multiple dimensions are not only more informative than single-item SRIs that ask an overall question such as “how effective was this instructor?” but they also facilitate critical analysis of validity evidence. Indeed, some relationships detract from the validity of SRIs with single-item scales, but they provide supporting validity evidence with SRIs that tap into multiple dimensions. Despite the challenges in summarizing the literature,

several authors have conducted comprehensive reviews and identified several important trends.

Aleamoni (1999) lists 16 claims that are often cited as reasons to avoid using SRIs. He then explains through a review of 154 studies that many of these claims are myths because of research that shows no support or inconclusive support for them. For example, Aleamoni cited 21 studies that either countered or failed to support the claim that SRIs are a popularity contest. He concluded that students were indeed able to distinguish between instructor friendliness and instructional characteristics such as course organization and student motivation.

The relationship between student ratings and class size is another example of a myth, but one that may not be so easily dismissed. Ten studies supported the claim that smaller classes result in higher SRI scores, but ten others did not (Aleamoni 1999). Although this evidence seems inconclusive, Aleamoni reported that an additional six studies showed evidence of a curvilinear relationship such that student ratings decreased as class size increased, but after a point they increased as class size increased. Marsh and Roche (2000) explained that this curvilinear relationship occurs with overall measures of teaching, but a linear relationship occurs in expected ways with multidimensional measures. Specifically, class size shows a small negative relationship with ratings of group interaction and student rapport, but no relationship with ratings of course coverage and organization. Given that an instructor's time to interact with every student in a large class is very limited, it makes sense for ratings of group interactions and rapport to be lower for large classes. It also makes sense that class size would be unrelated to other dimensions such as course organization. Thus, the effect of class size and whether it is validity evidence or bias may depend on the way teaching quality is measured.

Aleamoni (1999) classified student motivation and the grade/SRI relationship as myths, but his argument and the cited research is not convincing. Benton and Cashin's (2014) more recent review of the literature largely agrees with Aleamoni's review, but they acknowledged that these two characteristics are among those that may need to be controlled when evaluating teaching effectiveness in a summative way. The literature is quite clear that a relationship between course grades and SRI scores exists and is small to moderate in size (Aleamoni 1999; Clayson 2009; Cohen 1981; Johnson 2003; Marsh 1987). The real question is whether this small to moderate relationship represents a bias among respondents or favorable validity evidence. According to the grading leniency hypothesis, this relationship is due to students giving high ratings to instructors who give higher than deserved grades. This hypothesis is frequently encountered in the SRI literature as a reason to not use SRIs (Johnson 2003; Greenwald and Gillmore 1997). Johnson (2003) actually considers it to be the root cause of grade inflation in higher education.

The validity hypothesis is an alternative explanation for the grade/SRI relationship. According to this hypothesis, SRI scores and student achievement are influenced by teaching quality. That is, good teaching produces a large amount of learning as indicated by high grades, and students acknowledge good teaching with high ratings of instruction. The opposite effect is also true. Low teaching quality produces little learning and low grades, and students protest with low ratings of instruction. Marsh writes that multisection validity studies provide evidence supporting the validity hypothesis (Marsh 1987; Marsh and Roche 1997, 2000). These studies control for grading leniency through standardized course content and grading practices, yet they still result in moderate correlations between achievement and SRI scores. Cohen's (1981) meta-analysis of multisection validity studies reported an average correlation of 0.43 for overall instruction and 0.47 for overall course

quality. The average correlation with instructor skills was even higher ( $r = 0.50$ ). Clayson's (2009) more recent meta-analysis also showed small but positive correlations.

Although meta-analytic reviews of multisection validity studies support the validity hypothesis, recent studies have used value-added modeling to more directly examine the relationship between teaching quality and student learning. This methodology is an alternative to SRIs that attempts to isolate a teacher's contribution to student learning by controlling confounding variables such as prior achievement and student background characteristics (see McCaffrey et al. 2003). Carrell and West (2010) and later Braga et al. (2014) applied value-added modeling to a sequence of university-level courses to estimate a teacher's effect on learning in the course taught by the instructor (i.e., the contemporaneous course) and a teacher's effect on learning in a subsequent course that may have been taught by someone else (i.e., the follow-on course). They found that teachers who produced the most learning in the contemporaneous course were distinct from instructors who produced the most learning in a follow-on course. The researchers hypothesized that instructors who "teach to the test" have students that perform well in the contemporaneous course, but instructors who do not "teach to the test" promote understanding and ultimately create learning in the follow-on course. Carrell and West referred to this second type of learning as "deep learning." Complicating the matter, they also found that SRI scores were positively related to learning in the contemporaneous course, but negatively related to learning in a follow-on course. They argued that instructors who "teach to the test" received better SRI scores than instructor who produced "deep learning" among students. One takeaway is that if student learning in a follow-on course is the desired outcome, then SRIs are not a useful way to measure teaching quality. However, if learning in the course taught by the instructor is the preferred outcome, then SRIs are an acceptable way to measure teaching quality.

A limitation of the work by Carrell and West (2010) and Braga et al. (2014) is the design of their SRIs. Their measures involved single-item indicators that focused on a limited number of teaching quality dimensions. Students rated course organization, their perceived learning, and the overall course quality, but they did not rate the instructor's assessment practices or the academic rigor of the course. Omission of the later dimension is particularly noteworthy for two reasons. First, student ratings of academic rigor significantly predict teacher value-added scores, but student ratings of other dimension do not (Raudenbush and Jean 2014). Second, academic rigor is antithetical to "teaching to the test." It entails critical thinking, meta-cognitive strategies, and other practices that promote a deep understanding of course content. Had Carrell and West and Braga et al. included academic rigor in their SRIs, it would have been possible to determine if their "teaching to the test" hypothesis was true. If student ratings of academic rigor were negatively related to learning in the current course, but positively related to learning in a subsequent course, it would support the "teaching to the test" hypothesis. However, if student ratings of academic rigor were positively related to learning in the current course and a follow-on course, it would raise questions about the hypothesis. It would also raise questions about what is really being measured by value-added scores for a follow-on course.

In summary, our teaching framework and method of analysis are aimed at producing criterion-referenced score interpretation. However, the potential for bias from factors unrelated to teaching is real and should not be impetuously disregarded. Multidimensional SRIs that use multiple items per scale appear to be robust to these factors, but even quality measures should undergo careful review to evaluate the extent to which student and course characteristics influence scores.

## Methods

Our SRI originally<sup>2</sup> involved 24 items and a six point Likert scale that ranged from Strongly Disagree to Strongly Agree. The original theoretical framework and initial pool of items underwent expert review by five faculty from a university-wide teaching support center. Data collected from 106 volunteers from a participant pool during fall semester 2013 allowed us to conduct an item analysis, but the sample size was too small to evaluate the factor structure of the measure. Therefore, we used all of the items to measure a single dimension. Results showed that the measure had a coefficient alpha reliability estimate of 0.96. Results also indicated that few people selected the lowest response category. As a result, we limited the rating scale to five categories and changed the labels to range from Poor to Excellent. Question wording was revised to align it with the new response labels.

A second pilot study with the revised measure involved 170 volunteers. The coefficient alpha estimate of score reliability was 0.95. Item response theory analysis (details not shown) revealed that examinee scores were located higher on the latent trait than most step parameter estimates. This result means that our scale had less measurement precision in the region of the scale where there were many examinees. To improve measurement precision at the high end of the scale, we decided to write new items targeted at the high end of the scale to improve measurement precision at the high end. That is, we wrote items for which we thought high ratings would be rare. The revised measure included 30 items. This paper concerns a third pilot study that involved the 30-item measure.

## Participants

All participants were from a Mid-Atlantic university classified as an R1: Doctoral University—Highest Research Activity that enrolls about 15,000 undergraduate students and about 6000 graduate students. All courses in this study were offered by a school of education. Instructors represented all departments within the school, but not in equal proportions. A majority (44 %) represented a department focused on curriculum and instruction. The next largest group of instructors (37 %) represented a department focused on research and policy. Smaller percentages of instructors represented kinesiology (12 %), and human services (7 %). Most instructors were Associate Professors (41 %), but Assistant Professors (32 %), and Professors (26 %) were also represented in the sample. Fifty-six percent of instructors were female.

A total of 430 students from 65 different courses responded to the measure at the end of spring semester 2015. The percentage of students taking courses from each department was different from the percentage of instructors in each department. Students were enrolled in research and policy courses (37 %), kinesiology courses (32 %), curriculum and instruction courses (27 %), and human services courses (5 %). Courses varied in size with enrollments ranging from five to 160 students. The median enrollment was 26 students. A large portion of the courses (77 %) involved in-person only classes with fewer courses using only online classes (11 %) or a combination of in-person and online classes (12 %). Most students were seeking a Bachelor's degree (47 %) or Master's degree (34 %). Doctoral students (i.e., Ed.D., Ph.D.; 10 %), non-degree seeking (8 %), and professional students (i.e., M.D., J.D; 2 %) comprised smaller portions of the sample. Eighty percent of students were female. White students were the largest race/ethnicity group (80 %). Asian

<sup>2</sup> An unpublished manuscript about the original study is available upon request.



(8 %), Black/African–American (4 %), and Multiracial students (4 %) were the largest minority groups, but Hispanic (3 %), Native American/Alaskan Native (<1 %), and Other students (1 %) were also included in the sample.

## Procedures

The item response theory (IRT) analysis involved fitting Masters' (1982) partial credit model to data for each item. For an item indexed by  $j$  with  $m + 1$  response categories with ordinal scores  $k = 0, \dots, m$ , this model provides the probability that an examinee will select category  $k$ ,

$$P_{jk}(\theta) = \frac{\exp \sum_{l=0}^k (\theta - \delta_j - \tau_{jl})}{\sum_{s=0}^m \exp \sum_{l=0}^s (\theta - \delta_j - \tau_{jl})}.$$

It involves a person trait parameter,  $\theta$ , an item difficulty parameter,  $\delta_j$ , and  $m$  threshold parameters,  $\tau_{jl}$ . These parameters are on a logit scale that theoretically ranges from negative infinity to positive infinity. In practice, parameter estimates tend to range from  $-6$  to  $6$ . Threshold parameters indicate the intersection point, relative to item difficulty, for two adjacent characteristic curves. Note that IRT models have scales that are indeterminant up to a linear transformation, which means that constraints must be applied to identify the model. We followed the convention in Rasch measurement to constrain the mean item difficulty to zero, and the within-item threshold mean to zero. After estimating person and item parameters, we conducted principal components analysis of standardized residuals to evaluate dimensionality, implemented maximum information Wright mapping, and created anchor point descriptions to describe the latent scale.

## Wright Maps and Scale Anchor Point Descriptions

In defining criterion-referenced tests, Glaser (1963) recognized that achievement is defined on a continuum and the standard of performance is “the behavior which defines each point along the achievement continuum” and that “specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels” (p. 519). His writing was about student achievement, but it applies equally well to teaching quality. A Wright map<sup>3</sup> (Wilson 2005) is a method for implementing Glaser's concept of criterion-referenced score interpretation because it locates examinee scores and item response categories on the same latent continuum. As a result, item response categories at a particular location on the continuum can be used to describe performance at that point.

Item category locations may be identified in a number of ways such as mapping parameter estimate directly or choosing locations that result in a predefined response probability. We used a maximum information approach to item category mapping (Huynh 1998; Huynh and Meyer 2003). This method locates an item category at the place where it contributes the most information toward estimating the latent trait. Specifically, the maximum information location is the value of the latent trait,  $\theta$ , that maximizes the category information function  $I_{jk}(\theta) = P_{jk}(\theta)I_j(\theta)$ , where  $P_k(\theta)$  is the partial credit model for response category  $k$  and  $I_j(\theta)$  is the item information function.

<sup>3</sup> A Wright map is also referred to as an item map.

A Wright map shows the location of every response category for every item. Response locations can be consolidated into fewer scale anchor point descriptions by selecting anchor point locations and choosing item categories located at or below each anchor point (Huynh and Meyer 2003). The item text and response categories are then summarized into a qualitative description of the anchor point.

## Software

We used jMetrik (Meyer 2014) for the IRT analysis, computing fit statistics, and evaluating dimensionality. To study the extent to which threats to validity influence SRI scores, we used the lme4 (Bates et al. 2015) package in R to fit multilevel models to the data for each scale, and the R package lmerTest (Kuznetsova et al. 2015) to test the significance of fixed effects.

## Results

### Missing Data

We scored each item on a zero to four point scale, where the lowest score was for a response of Poor and the highest score was for a response of Excellent. Students also had the option of choosing a response of Not Applicable for each item. The frequency of Not Applicable responses ranged from a low of two to a high of 55. Omitted responses were less common and occurred from a low of once to a high of five times. We treated Not Applicable and omitted responses as missing data during the analysis. For the classical item analysis and differential item functioning analysis, missing responses were scored as zero points. In the IRT analysis, missing data were ignored, and estimates made use of all available data.

### Dimensionality

We originally hypothesized that our construct formed a single dimension with multiple parts (e.g., Organization and structure, Academic rigor). However, principal components analysis of standardized residuals from the partial credit model (Linacre 2006) indicated the presence of multiple secondary dimensions (results not shown). Consequently, we divided the measure into four separate scales with each scale measuring a separate dimension. Items were exclusively assigned to one scale according to the factor loadings and question wording. We also elected to remove four items. Subsequent principal components analysis of standardized residuals indicated that each scale measured a single dimension. In particular, the eigen value for the first principal component of the standardized residuals was 1.74 for Organization and structure, 1.96 for Assessment and feedback, 1.83 for Personal interaction, and 1.78 for Academic rigor. These values are all less than two, which indicates that each scale taps into a single dimension (see Linacre 2006). Although data supported the use of four separate scales, they are all correlated above 0.75 (see Table 1). It is possible that together the scales measure a common dimension as well as each specific dimension (i.e., a bifactor or testlet model). However, we did not test that type of factor structure because our sample size was too small for such

**Table 1** Correlations and descriptive statistics for individual responses to each scale

Scale	Organization and structure (OS)	Assessment and feedback (AF)	Personal interactions (PI)	Academic rigor (AR)
OS	1.00			
AF	0.83	1.00		
PI	0.75	0.76	1.00	
AR	0.86	0.86	0.75	1.00
Raw score mean	19.16	18.39	13.59	27.67
Raw score SD	5.31	6.10	3.24	8.80
Reliability	0.89	0.90	0.80	0.92
SEM	1.73	1.95	1.73	2.51
Number of items	6	6	4	9
Scaled score mean	128.50	126.80	136.40	125.60
Scaled score SD	19.77	20.39	17.98	19.96

a complex model. In keeping with this new structure of the measure, we evaluated the psychometric quality of each scale separately.

### Reliability of Individual Scores and Course Mean Scores

Prior to estimating score reliability, we ran a classical item analysis and differential item functioning analysis as an initial check on the quality of the TQSRI items. Details of that analysis are not included here, but all items met standards of quality in terms of classical item difficulty, item discrimination, and differential item functioning.

Table 1 lists estimates of reliability and the standard error of measurement (SEM) for individual scores. The Personal interaction scale had the lowest reliability estimate (0.80), albeit a good value for reliability. Organization and structure, Assessment and feedback, and Academic rigor had higher reliability estimates of 0.89, 0.90, and 0.92, respectively.

Given that decisions about teaching quality often focus on course means, we also computed the reliability of course means using variance components estimated in the multilevel models described below. This reliability estimate is given by  $r = \tau^2 / (\tau^2 + \sigma^2/n)$ , where  $\tau^2$  is the between course variance,  $\sigma^2$  is the within course variance, and  $n$  is the number of students in the course (see Lüdtke et al. 2009). Thus, each class has its own reliability estimate. The average reliability of course means is listed in Table 2. Personal interactions had the lowest reliability of course means with reliabilities ranging from 0.21 to 0.95. Assessment and feedback had the next lowest reliability of course means with values that ranged from 0.23 to 0.96. Next, Organization and structure had course mean reliabilities that ranged from 0.27 to 0.96. Finally, Academic rigor had the highest course mean reliability with values that ranged from 0.30 to 0.97.

Note that the reliability of course means increases as the number of students responding increases. That is, the smallest course mean reliability estimates were for courses with the lowest number of respondents, and the highest course mean reliability estimates occurred for the courses with the largest number of respondents. This relationship raises the question, “how many students must complete the SRI to have a reliable course mean?” By using our estimated variance components in the equation for course mean reliability and substituting different values for the course size, we project that the number of students

**Table 2** Correlations and descriptive statistics for scaled score course means

Statistic	Scale			
	Organization and structure (OS)	Assessment and feedback (AF)	Personal interactions (PI)	Academic rigor (AR)
OS	1.00			
AF	0.87	1.00		
PI	0.80	0.80	1.00	
AR	0.88	0.90	0.82	1.00
Min	94.34	89.60	97.33	84.22
Max	150.00	150.00	150.00	150.00
Mean	128.30	128.40	137.00	126.80
SD	14.53	14.56	12.60	15.05
ICC	0.27	0.23	0.21	0.30
Reliability	0.61	0.56	0.53	0.64

needed for a course mean reliability of at least 0.7 is seven for Organization and structure, eight for Assessment and feedback, nine for Personal interactions, and six for academic rigor. However, these projections depend on the estimated variance components and are somewhat sample dependent much like the Spearman–Brown formula.

### Item Response Theory and Test Scaling

Table 3 lists partial credit model item parameter estimates for every item on each scale. Item difficulty estimates ranged from  $-0.42$  to  $0.43$  for Organization and structure, from  $-0.56$  to  $0.72$  for Assessment and feedback, from  $-1.17$  to  $0.73$  for Personal interactions, and from  $-0.48$  to  $0.27$  for Academic rigor. Threshold estimates were properly ordered for every item. For example, the threshold between poor and fair,  $\tau_{j1}$ , was the lowest, the threshold between fair and good,  $\tau_{j2}$ , was the next lowest, and so on until the highest threshold estimate for very good and excellent,  $\tau_{j4}$ .

Bond and Fox (2007) note that infit and outfit value between 0.6 and 1.4 indicate good fit for the partial credit model. Almost all of our infit and outfit values were within this recommended range. Two items had infit and outfit values slightly above 1.4 (see Table 3), but these values were not large enough to warrant removal of the items.

For each scale, we linearly transformed person trait estimates to scaled scores that ranged from 50 to 150. Details of this transformation are not provided, but Table 1 lists scaled score descriptive statistics. On average scores are near the upper end of the scale, which indicates that students tended to provide high ratings of instruction. All four scales are correlated between 0.75 and 0.86 (see Table 1). Organization and structure, Assessment and feedback, and Academic rigor have notably higher correlations with each other than they do with Personal interactions. Descriptive statistics for course means have a similar pattern to statistics for individual scores (see Table 2). Correlations among scales are slightly larger for course means than they were for individual scores.

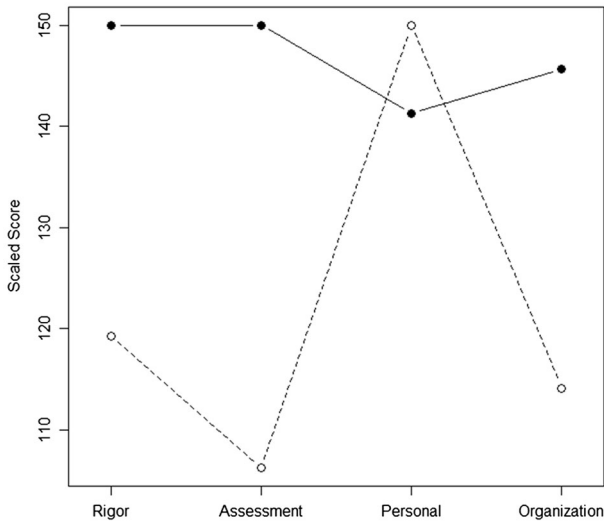
**Table 3** Partial credit model item difficulty and threshold parameter estimates and item fit statistics

Scale	Item	$\delta_j$	$\tau_{j1}$	$\tau_{j2}$	$\tau_{j3}$	$\tau_{j4}$	Infit	Outfit
Organization and structure	OMFA14001	-0.33	-3.14	-1.36	1.28	3.22	0.87	0.90
	OMFA14003	-0.42	-2.54	-1.64	1.29	2.89	0.81	0.76
	OMFA14006	0.01	-3.75	-0.32	1.17	2.90	1.07	1.13
	DPFA14001	0.12	-3.51	-0.87	1.24	3.14	0.91	0.91
	DPFA14006	0.18	-2.70	-1.25	0.95	3.00	0.92	0.92
	AEFA14001	0.43	-3.34	-0.89	1.22	3.01	1.30	1.30
Assessment and feedback	AEFA14002	-0.09	-2.07	-1.19	0.72	2.55	0.88	0.90
	AEFA14003	0.72	-2.47	-1.14	1.28	2.33	0.84	0.81
	AEFA14004	0.35	-2.79	-0.81	0.91	2.69	1.45	1.43
	AEFA14005	-0.56	-3.26	-0.41	0.96	2.71	0.84	0.80
	AEFA14006	-0.24	-2.17	-0.86	0.66	2.37	0.65	0.66
	ARSP15001	-0.18	-2.03	-1.12	0.22	2.94	1.11	1.28
Personal interactions	OMFA14004	0.52	-3.03	-1.01	0.67	3.37	1.10	1.09
	OMFA14005	0.20	-4.11	-0.60	0.85	3.86	1.11	1.13
	PIFA14001	0.73	-2.79	-1.11	0.86	3.04	0.92	0.98
	PIFA14003	-1.17	-2.39	-1.20	0.21	3.38	1.03	0.96
	PIFA14004	-0.28	-4.02	-0.21	0.87	3.36	0.76	0.78
Academic rigor	OMSP15007	0.27	-2.06	-1.05	0.63	2.48	0.92	0.91
	OMSP15008	0.08	-2.85	-0.55	0.91	2.50	1.05	1.00
	DPFA14002	0.20	-2.97	-0.73	0.84	2.86	0.97	0.97
	DPFA14003	-0.31	-2.56	-1.19	1.04	2.71	0.75	0.73
	DPFA14005	-0.48	-2.53	-0.59	1.05	2.07	0.92	0.95
	DPFA14008	0.25	-2.66	-0.63	0.74	2.55	1.42	1.42
	DPFA14009	-0.24	-2.43	-1.20	0.95	2.68	1.17	1.17
	PIFA14002	0.03	-2.29	-0.34	0.76	1.87	0.71	0.70
	ARSP15002	0.19	-2.19	-0.41	0.24	2.36	1.03	1.16

Parameters estimated separately for each scale. Fit was computed within each scale

## Score Interpretation

Score profiles provide a way to identify an instructor's strengths and areas that need more development. For example, Fig. 1 illustrates the score profile for two courses. One instructor scored high on all scales, but the instructor's lowest score was for Personal interactions. This profile contrasts with the scores for the second instructor who scored high on Personal interactions (higher than the other instructor), but low on Organization and structure, Assessment and feedback, and Academic rigor. It is clear from the profiles that these instructors have different needs. One could benefit from better communication and interaction with students, while the other could benefit by improving the structure, assessment, and rigor of the course. Score profiles use multiple dimensions to paint a broad picture of instructional quality, but they alone do not result in criterion-referenced score interpretation.



**Fig. 1** Scaled score profiles for two courses

To create a criterion-referenced interpretation of scores, we produced Wright maps and consolidated the information into four evenly spaced anchor point descriptions for each scale (see Tables 4, 5, 6, 7). Each anchor point description involved three to four item categories. Italicized text in each description indicates the response category. The best anchor point description of student ratings for a course is the one that is closest to, but not larger than, the course mean. For example, if an instructor had an average score of 101 on Academic rigor, the anchor point description for a score of 96 in Table 7 would best describe student ratings for this instructor.

Wright maps and scale anchor point descriptions also provides guidance on how to refine a course or improve pedagogy because characteristics located above the course mean are unlikely to be selected by students. For example, the description for an academic rigor score of 115 is uncharacteristic of the teacher with a course mean of 100, but it involves the characteristics that the teacher should aim to improve in order to improve mean performance. That is, the teacher with a mean score of 100 should make efforts to help students to think about the course content in a structured and organized way and help students make connections to content and skills learned in other courses.

## Evaluation of Threats to Validity

Multilevel models account for students being nested within courses, and allow the study of student- and course-level covariates that may affect student ratings. Student-level variables included whether or not the course was required or elective (REQUIRED), whether or not the student was angry at the instructor for any reason (ANGRY), the student's degree program (i.e., Bachelor's, Master's, Doctoral, Professional, or None), the student's type of enrollment (i.e., Full-time, part-time, or nondegree-seeking), the student's minority status (MINORITY), and whether or not the student was female (FEMALE). We named variables in a manner consistent with their dummy coding. For example, the variable FEMALE was dummy coded such that female = 1 and male = 0. Similarly, the variable ANGRY was

**Table 4** Organization and structure scale anchor point descriptions

Score	Anchor point description
134	Students rate the instructors as <i>Excellent</i> at explaining grading procedures, providing organized course materials, and presenting course content in an organized fashion.
115	Students rate the instructor as <i>Very Good</i> at adhering to course learning objectives, but <i>Good</i> at explaining grading procedures, and <i>Good</i> at clearly stating requirements and due dates for assignments.
97	Students rate the instructor as <i>Good</i> at Adhering to course learning objectives for the duration of the course, but only <i>Fair</i> at clearly stating requirements and due dates for assignments and explaining grading procedures.
79	Students rate the instructor as <i>Poor</i> at providing organized course materials, defining learning objectives for the course, and explaining grading procedures.

dummy coded as 1 if the student indicated he or she was angry at the instructor and 0 otherwise. Course-level variables included whether or not the instructor was female (INSFEMALE), the number of students enrolled in a course (SIZE), and the SRI response rate (RESPONSE). The latter two variables were grand mean centered. Using student scores from one of the evaluation scales (SCORE) as the dependent variable, the level-1 model is given as follows

$$\begin{aligned}
 \text{SCORE} = & \beta_{0j} + \beta_{1j}(\text{REQUIRED}) + \beta_{2j}(\text{ANGRY}) + \beta_{3j}(\text{MASTERS}) \\
 & + \beta_{4j}(\text{DOCTORAL}) + \beta_{5j}(\text{PROFESSIONAL}) + \beta_{6j}(\text{NONE}) \\
 & + \beta_{7j}(\text{PARTTIME}) + \beta_{8j}(\text{NONDEGREE}) + \beta_{9j}(\text{MINORITY}) \\
 & + \beta_{10j}(\text{FEMALE}) + r_{ij}.
 \end{aligned}$$

The level-2 model for the intercept is given by,

$$\begin{aligned}
 \beta_{0j} = & \gamma_{00} + \gamma_{01}(\text{SIZE}) + \gamma_{02}(\text{RESPONSE}) + \gamma_{03}(\text{INSTFEMALE}) + \gamma_{04}(\text{FEMALE} \\
 & \times \text{INSTFEMALE}) + u_{0j}.
 \end{aligned}$$

For each scale, we fit an unconditional model, a model with student-level covariates, and a model with student- and course-level covariates. Using results from the unconditional model, intraclass correlations for each outcome indicated that course accounted for 20–30 % of the total variance (see Table 2). Academic rigor had the largest ICC. Reliability estimates for course means were reported earlier. Adding student-level covariates explained 9 % of student-level variance for Organization and structure, 7 % of student-level variance for Assessment and feedback, 12 % of student-level variance for Personal interactions, and 11 % of student-level variance for academic rigor. Adding course-level covariates significantly improved model fit for every scale. The Chi square value for the

**Table 5** Assessment and feedback scale anchor point descriptions

Score	Anchor point description
136	Students rate the instructor as <i>Excellent</i> at giving feedback on student progress throughout the semester. Moreover, this feedback is timely and occurs frequently throughout the semester. Students also rate the instructor as <i>Excellent</i> at requiring students to demonstrate mastery of course content or skills.
116	Students rate the instructor as <i>Very Good</i> at giving assignments or exams that were relevant to course learning objectives and providing opportunities to apply skills and learning to real-world problems. Moreover, student rated the instructor as <i>Very Good</i> at giving assignments that challenged students to think critically.
96	Students rate the instructor as <i>Good</i> at giving assignments or exams that were relevant to course learning objectives and requiring students to demonstrate mastery of content and skills taught in class. However, students only rated the instructor as <i>Fair</i> at giving students feedback on their progress throughout the semester.
76	Students rate the instructor as <i>Poor</i> at providing assignment or exams that challenged them to think critically, <i>Poor</i> at grading assignments and providing timely feedback, and <i>Poor</i> at providing opportunities to apply skills learned in class to real-world problems.

comparison of the student-covariate model to the full model was 19.965 ( $p < 0.001$ ) for Organization and structure, 18.823 ( $p < 0.001$ ) for Assessment and feedback, 20.358 ( $p < 0.001$ ) for Personal interactions, and 20.974 ( $p < 0.001$ ) for Academic rigor.

Table 8 lists fixed and random effect estimates for each model. ANGRY is statistically significant at the 0.001 level in every analysis. Angry students represented 4 % of the sample and 10 courses had one student identify as being angry, one course had three students identify as angry, and another course had four student identify as angry at the instructor. Despite the low prevalence in the sample, angry students rate instructors an average of 29.27 points lower on Organization and structure, 28.79 points lower on Assessment and feedback, 26.62 points lower on Personal interactions, and 28.25 points lower on Academic rigor.

The fixed effect for PARTTIME was statistically significant at the 0.05 level in the analysis of Organization and structure and Assessment and feedback such that part-time students rate instructors higher than full-time students. Part-time students also rated instructors higher than full-time students on Personal interactions and academic rigor, but these effects were not statistically significant.

Among the level-2 predictors, class size showed small positive relationship with scores, and the response rate showed a small negative relationship. The patterns of these relationships were consistent across all scales, but they were not statistically significant. The interaction between student gender (FEMALE) and instructor gender (INSFEMALE) was the only significant effect for level-2 covariates. Figure 2 illustrates the interaction for each scale using effects based on the multilevel model (i.e., nesting and lower-order terms).



**Table 6** Personal interactions scale anchor point descriptions

Score	Anchor point description
132	Students rate the instructor as <i>Excellent</i> at being available outside the classroom to provide help with the course, <i>Excellent</i> at responding to communications in a timely manner, and <i>Excellent</i> at demonstrating concern for student learning.
114	Students rate the instructor as <i>Very Good</i> at being approachable for questions or conversation about the course, but only <i>Good</i> at demonstrating concern for student learning, and <i>Good</i> at being available outside the classroom to provide help with the course.
96	Students rate the instructor as <i>Fair</i> at being approachable for questions or conversations about the course, being available outside the classroom to provide help with the course, and demonstrating concern for student learning.
79	Students rate the instructor as <i>Fair</i> as treating students fairly and without prejudice toward race, gender and other personal characteristics. However, students rate the instructor as <i>Poor</i> at responding to communications in a timely manner, and <i>Poor</i> at being available outside the classroom for help with the course.

Male students tend to rate faculty the same way, regardless of gender. However, female student rate female instructors higher than male instructors.

The full model for each scale fits the data best, but course-level covariates only account for 8 % of the course level variance for Organization and structure, 25 % of course-level variance for Assessment and feedback, and 12 % for Academic rigor. Course-level variance for Personal Interactions actually increased by 15 % with the addition of course-level covariates. Thus, the significant course-level covariates are relatively weak effects.

## Discussion

The Teaching Quality Student Rating Inventory is a criterion-referenced measure of teaching that involves Organization and structure, Assessment and feedback, Personal interactions, and Academic rigor. We created a Wright map for each scale and consolidated the information into anchor point descriptions. These maps and descriptions provide a criterion-referenced interpretive framework for scores by making a direct link between teaching characteristics and test scores. Although these maps aid score interpretation, research suggests several potential sources of bias in SRI scores.

Our multilevel analysis identified two possible threats to validity. Student anger and the interaction between student gender and instructor gender were statistically significant for all scales. These effects accounted for a relatively small amount of variance in student- and class-level scores, but we looked further into potential explanations of these effects. The

**Table 7** Academic rigor scale anchor point descriptions

Score	Anchor point description
134	Students rate the instructor as <i>Excellent</i> at explaining concepts in a clear and understandable way, <i>Excellent</i> at making connections to content and skills learned in other courses, and <i>Excellent</i> at integrating information technology into instruction.
115	Students rate the instructor as <i>Very Good</i> at helping students think about course content in a structured and organized way, and <i>Very Good</i> at making connections to content and skills learned in other course. Students also rate the instructor as <i>Very Good</i> at integrating information technology into instruction.
96	Students rate the instructor as <i>Good</i> at making an effort to address issues of diversity and equity as it relates to the course. However, student only rate the instructor as <i>Fair</i> at stimulating students' intellectual curiosity and <i>Fair</i> at inspiring and motivating student interest in course content.
77	Students rate the instructor as <i>Poor</i> at explaining concepts in a clear and understandable way and challenging students to think critically about course content. They also rate the instructor as <i>Poor</i> at integrating information technology into instruction, and <i>Poor</i> at addressing issues of diversity and equity as it relates to the course.

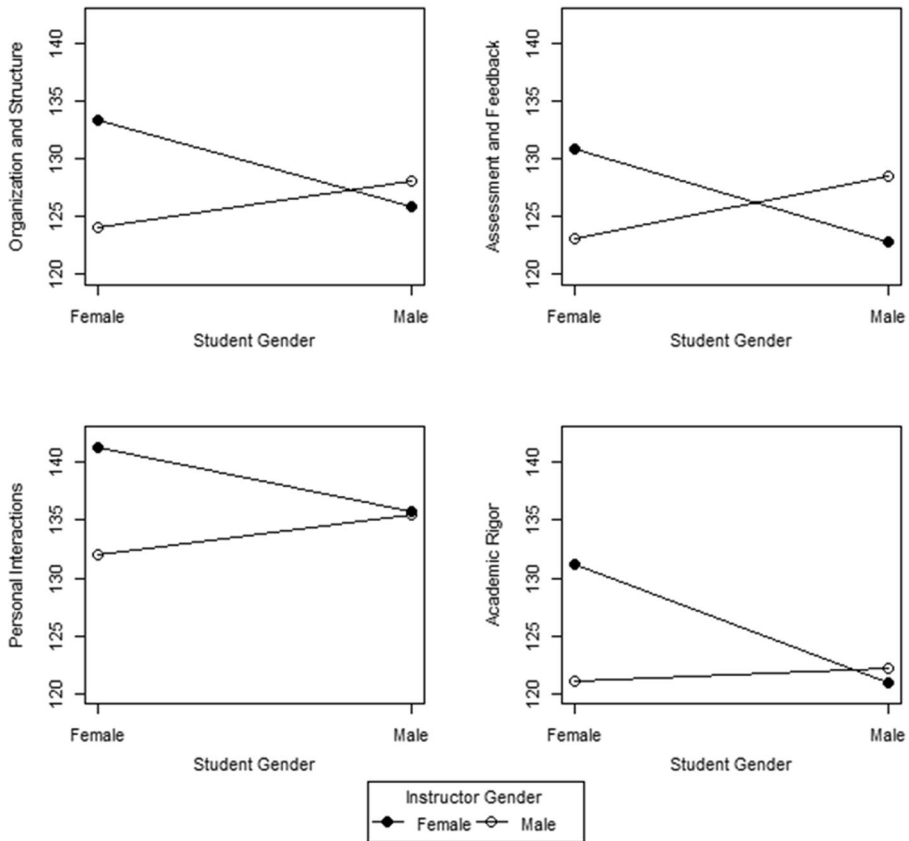
question about student anger simply asked if the student was angry at the instructor for any reason. We did not ask for more specific information. Students may identify as angry because of the low quality of the course. In this case, the association between anger and low ratings is aligned with the use of the scales (i.e., evidence in favor of test validity). However, students may also identify as angry because they are upset about a low course grade. Low ratings by these angry students may be vindictive responses and not an honest critique of teaching quality (i.e., evidence against test validity). We explored the possibility of these two explanations by ranking course evaluation mean scores and summarizing the ranks for courses that had at least one student identify as being angry (details not shown). One course with multiple students indicating anger at the instructor was consistently ranked among the lowest courses. This suggests that these students are angry about the low quality of the course. However, the picture is not entirely clear. A few courses with an angry student were ranked among the highest courses, and a couple of the lowest ranked courses had no students identify as being angry; the implication is that some students are angry about their grade or some other factor unrelated to teaching quality. Taken together, angry students clearly rate instructors lower on all scales, but the source of their anger is unclear. Future research is needed to study the source of anger in more detail.

We turned to prior research to further understand the significant gender interaction. Basow (1995) also found significant but small interactions between instructor gender and student gender. She noted the same-gender preference was stronger for females students,

**Table 8** Fixed and random effects for all scales

Fixed effects	Organization and structure			Assessment and feedback			Personal interactions			Academic rigor		
	Estimate	SE	df	Estimate	SE	df	Estimate	SE	df	Estimate	SE	df
Intercept	128.81 <sup>†</sup>	4.77	184.6	127.22 <sup>†</sup>	4.70	184.0	133.72 <sup>†</sup>	4.29	179.7	122.64 <sup>†</sup>	4.76	177.8
REQUIRED	-1.29	2.48	392.3	1.03	2.53	377.1	0.55	2.24	390.0	1.44	2.45	394.5
ANGRY	-29.27 <sup>†</sup>	4.57	384.7	-28.79 <sup>†</sup>	4.73	391.5	-26.62 <sup>†</sup>	4.15	385.3	-28.25 <sup>†</sup>	4.49	381.7
MASTERS	0.55	3.02	376.4	0.41	3.07	349.5	3.63	2.73	371.3	0.19	2.98	381.1
DOCTORAL	3.01	4.10	370.7	6.70	4.15	329.6	6.27	3.70	363.3	-0.05	4.05	377.7
PROFESSIONAL	-1.98	5.90	374.2	0.15	6.13	382.1	5.96	5.35	374.3	1.42	5.80	370.6
NONE	-9.01	8.07	395.5	-5.25	8.32	397.9	-6.27	7.32	396.0	-3.23	7.94	394.1
PARTTIME	8.57*	3.76	245.9	9.12*	3.74	213.0	2.69	3.39	234.9	6.97	3.74	250.5
NONDEGREE	12.83	7.76	395.1	14.11	8.00	397.4	9.48	7.03	395.4	9.51	7.63	394.0
MINORITY	2.24	7.76	377.1	3.31	2.23	381.7	3.79	1.95	376.9	3.03	2.11	374.5
FEMALE	-3.97	3.70	379.4	-5.35	3.83	387.7	-3.33	3.35	380.0	-1.07	3.63	375.8
INFEMALE	-2.29	5.34	176.6	-5.76	5.26	185.5	0.26	4.80	173.3	-1.27	5.34	167.6
FEMALE × INFEMALE	11.51*	4.59	385.6	13.26**	4.75	392.4	8.85*	4.16	386.3	11.22*	4.51	382.8
SIZE	0.10	0.06	33.0	0.04	0.06	26.8	0.08	0.06	30.4	0.09	0.07	33.0
RESPONSE	-0.09	0.09	44.9	-0.09	0.08	37.5	-0.06	0.08	41.7	-0.15	0.09	44.5
Random effects	Variance	SD	Variance	SD	Variance	SD	Variance	SD	Variance	SD	Variance	SD
Course	103.70	10.18	73.50	8.57	79.51	8.92	112.00	10.58				
Residual	272.00	16.49	298.00	17.26	224.41	14.98	261.10	16.16				
Model deviance	3538.5		3560.6		3456.3		3547.3					

\* p < 0.05; \*\* p < 0.01; <sup>†</sup> p < 0.001



**Fig. 2** Gender interaction plots for each scale

especially on questions related to feedback, respect for students, and freedom to express ideas. Anderson and Miller (1997) explain that this interaction is due to different teaching styles among male and female faculty and different expectations for teaching by male and female students. Female-oriented teaching characteristics described by Basow and Anderson and Miller are very similar to those included on our Assessment and Feedback and Personal Interactions scales where the course-level covariates, including the student-faculty gender interaction, had stronger effects than on other scales. Thus, our results comport with others who conclude that gender effects are small or inconsistent (Aleamoni 1999; Anderson and Miller 1997; Basow 1995; Benton and Cashin 2014; Feldman 1992, 1993).

In summary, validity evidence and score reliability supports the idea that the TQSRI is a suitable measures for making instructional decisions and refining a course. It and other SRIs should be used more cautiously when making high-stakes personnel decisions such as promotion and tenure. One reason for this caveat is that course means are the basis of comparing one instructor's performance to another's, yet the reliability of course means can be low when only a few students complete the ratings. Another reason is the small but statistically significant effects of gender and student anger that was found in all four scales. If these effects are too much of a concern for making high-stakes decisions, then they and

other potential sources of bias could be statistically controlled (Benton and Cashin 2014; Greenwald and Gillmore 1997; Johnson 2003) or results could be disaggregated by gender or other demographics.

## Limitations

The main limitation is the lack of data on student performance. We did not ask students to report their actual or perceived grade in the course. Students completed the survey anonymously and there was no way to link students with existing data on student achievement such as GPA. Consequently, we were unable to examine the relationship between grades and SRI scores.

In the future we recommend, for ourselves and others, that SRIs include questions about grading practices and instructor characteristics to help further evaluate the grade/SRI relationship. Grading practice questions could ask whether the grading was “an easy A” or “more strict than deserved.” We also recommend the use of data about instructor characteristics such as past teaching awards, years teaching, and hours of professional development related to teaching. These instructor characteristic would provide evidence of validity, if they were related to SRI scores. We were unable to collect this information, but these variables would provide more information about the validity of SRI scores.

Our data were also limited to courses taught at one school within the university. It is possible that results from our validity study may be specific to this school and may change in direction or magnitude with data from across the university. Basow (1995) found that the student and instructor gender interaction differed by type of school (i.e., humanities, social sciences, physical sciences). Therefore, future research might seek to include students across various disciplines and schools and study the way validity results might vary by school or discipline.

**Acknowledgments** We thank Emily Bowling, Fares Karam, Bo Odom, and Laura Tortorelli for their work on the original version of this measure. They developed the original teaching framework and wrote the initial pool of items as part of a course project.

## References

- Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153–166.
- Anderson, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics*, 30(2), 216–219.
- Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems* (3rd ed.). Bolton, MA: Anker Publishing Company Inc.
- Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, 87(4), 656–665.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.
- Bensley, D. A. (2010). A brief guide for teaching and assessing critical thinking in psychology. *Observer*, 23(10). Retrieved from <http://www.psychologicalscience.org/index.php/publications/observer/2010/december-10/a-brief-guide-for-teaching-and-assessing-critical-thinking-in-psychology.html>.
- Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues & Practice*, 17, 10–22.

- Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory & research* (Vol. 29, pp. 279–326). Dordrecht, The Netherlands: Springer.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.
- Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economic of Education Review*, *41*, 71–88.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Brophy, J. E. (1999). *Teaching (educational practices series—1)*. Geneva, Switzerland: International Academy of Education and International Bureau of Education, UNESCO.
- Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), 409–432.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards for tests*. Thousand Oaks, CA: Sage.
- Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*, 445–459. doi:10.3102/00346543053004445.
- Clark, R. E. (2009). Translating research into new instructional technologies for higher education: The active ingredient process. *The Journal of Computing in Higher Education*, *21*, 4–18. doi:10.1007/s12528-009-9013-8.
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, 16–30.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281–309.
- Feldman, K. A. (1992). College students' views of male and female college teachers: Part I—evidence from the social laboratory and experiments. *Research in Higher Education*, *33*(3), 317–375.
- Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211.
- Ferguson, R. F. (2012). Can student surveys measure teaching quality. *The Phi Delta Kappan*, *94*(3), 24–28.
- Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses*. San Francisco: Jossey-Bass.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, *18*, 519–521.
- Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*, 1209–1217.
- Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., et al. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4000 classrooms. *The Elementary School Journal*, *113*(4), 461–487.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, *23*, 35–56.
- Huynh, H., & Meyer, J. P. (2003). Maximum information approach to scale description for affective measures based on the Rasch model. *Journal of Applied Measurement*, *4*, 101–110.
- Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York: Springer.
- Kennedy, M. J., Thomas, C. N., Aronin, S., Newton, J. R., & Lloyd, J. W. (2014). Improving teacher candidate knowledge using content acquisition podcasts. *Computers & Education*, *70*, 116–127. doi:10.1016/j.compedu.2013.08.010.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest' [computer software] version 2.0–29. Retrieved from <https://cran.r-project.org/web/packages/lmerTest/index.html>.
- Linacre, J. M. (2006). *A user's guide to WINSTEPS Rasch-model computer programs*. Chicago, IL: Author.
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*, 120–131.
- 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., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187–1197.

- Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myths, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202–228.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.
- Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, *63*, 760–769. doi:10.1037/0003-066X.63.8.760.
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York: Cambridge University Press.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, D. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand.
- McKeachie, W. J., & Svinicki, M. (2006). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (12th ed.). Boston: Houghton Mifflin.
- Meyer, J. P. (2014). *Applied measurement with jMetrik*. New York: Routledge.
- Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *109*, 27–44. doi:10.1002/ir.2.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.
- Pyc, M. A., Agarwal, P. K., & Roediger, H. L., III (2014). Test-enhanced learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.) *Applying science of learning in education: Infusing psychological science into the curriculum*. Retrieved from the Society for the Teaching of Psychology website <http://teachpsych.org/ebooks/asle2014/index.php>.
- Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems*. San Francisco, CA: Jossey-Bass.
- Wiggins, G., & McTighe, J. (2011). *Understanding by design* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*. Washington, D.C.: American Federation of Teachers. <http://www.aft.org/newspubs/periodicals/ae>.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.