To: Faculty Senate

From: The Educational Policy Committee
      Cindy Baccar, Rowanna Carpenter, Enrique Cortez, Sri Craven, David Hansen (Ex-Officio), Steve Harmon (Consultant), Lisa Hawash, Art Hendricks (Co-Chair), Alison Heyer, Leslie McBride, John Ott, David Raffo (Co-Chair), Leopoldo Rodriguez, Alex Sager, Ken Stedman, Rolf Widenhorn, Hyeyoung Woo, Hormoz Zareh

Date: February 25, 2018

Subject: Draft: Use of Students Ratings of Instruction in Faculty Evaluation

1. Introduction

At Portland State University (PSU), Students’ Ratings of Instruction (SRI) (aka course evaluations) are often the primary means for evaluating faculty teaching for promotion & tenure, post-tenure reviews, as well as for non-tenure track faculty (NTTF) and adjunct faculty reviews. Given the prominence and weight of SRIs in these important decision processes, the Educational Policy Committee (EPC) of the Faculty Senate (FS) has reviewed some of the literature surrounding the best practices for the validity, construction and use of SRIs in evaluations of teaching quality. This report examines some of the key issues and makes recommendations pertaining to valid instrument construction as well as valid analysis and interpretation of the resulting SRI data.

In terms of general policy and consistent with best practice, the EPC sees the content of the SRI as the purview of academic departments. It is our intention that this report provides guidance and recommendations on ways to improve and strengthen PSU’s system of evaluation and thereby improves education quality and provides valid feedback for faculty and the university. It is also hoped that this report will be helpful in establishing the minimum requirements for any information system acquired by PSU/OIT to support the collection and analysis of SRI data. It is important that such a system support PSU’s ability to follow the best practices identified for creating valid SRI instruments. Finally, EPC hopes that the university, by implementing these described best practices, will reduce the potential legal liability associated with high stakes decisions being made using invalid instruments, by conducting invalid analyses, or by misinterpreting SRI data.

The EPC has reviewed research on the validity, reliability, and use of SRIs for the evaluation of faculty teaching to identify best practices for instrument construction, data collection, analysis, interpretation, and use of SRI. As part of this process, the EPC also consulted with multiple groups within Portland State, including the Office of Academic Innovation (OAI). This report presents the EPC’s observations and recommendations for the use of SRIs in high-stakes decisions regarding faculty teaching performance.

2. Background

Student ratings are the primary, and often the only, form of evidence available for the purpose of assessing teaching effectiveness as part of review and promotion processes. SRIs are a readily available source of evidence related to faculty teaching, which makes them a convenient and readily obtainable source of information about a faculty member’s teaching. While SRIs are an important source of information, they provide only one perspective – the student’s – on a faculty member’s teaching.
effectiveness. As noted below, students do not have the knowledge, perspective, training, or experience to judge the faculty member’s subject area knowledge or the currency of the course material. Other types of evidence are needed to supplement the information provided by SRIs.

This report focuses on recommendations for improving the validity of SRIs. We anticipate that a future report will focus on a larger set of recommendations related to the array of evidence that may be appropriate for evaluating overall teaching effectiveness.

### 3. **Best Practice Suggests that Multiple Sources of Assessment be Used**

Before reviewing the use of SRIs for evaluation of teaching, best practice strongly suggests that student ratings should be “only one source of data; they must be combined with additional evidence so that administrators and faculty peers can make an informed judgment about teaching quality.” (Hoyt and Pallett, 1999). Multiple sources of information – including student ratings – should be used to provide sufficient information to make a valid judgment about an instructor’s *overall teaching effectiveness* (Benton & Cashin, 2012).

### 4. **SRI Instrument Creation**

An SRI will only provide reliable, valid, and useful information to faculty and administrators if it is well constructed. *Our recommendations about analysis, use, and interpretation of SRI data are all based on the understanding that the data is derived from a well-constructed SRI instrument.*

A well-constructed SRI will recognize teaching as multi-dimensional and will focus questions on aspects of teaching that students are in a good position to rate. As a first step toward a well-constructed SRI, departments need to determine the specific dimensions, or constructs, they wish to measure, such as student satisfaction, effective teaching practices, or learning outcomes. Then, given those dimensions, SRI items can be identified or developed. Those items will include questions that serve both formative and summative evaluation purposes.

The EPC recommends that departments consult with a knowledgeable professional/researcher to assist them in creating an SRI – including the questions, the development of appropriate scales, as well as evaluating for validity and bias.

In order to conduct the most robust analyses, each construct should have several items (questions) associated with it. Items that are associated with each other in this way constitute a scale. A well-constructed Likert scale demonstrates the following characteristics:

1. Comprised of several (four to six) items (questions) (Hinkin et. al. 1997)
2. Horizontally arranged response levels
3. Response levels anchored with consecutive integers (e.g., 1, 2, 3, 4, 5)
4. Response levels verbally labeled to connote essentially evenly-spaced gradations that are bivalent and symmetrical to a neutral midpoint (e.g., strongly agree, agree, neutral, disagree, strongly disagree)
5. Measures attitude in terms of agreement/disagreement with a target statement.
Any SRI system that is adopted needs to be able to analyze and present data appropriately as well as collect important contextual variables. Failure to clearly identify the constructs of effective teaching and use of Likert scales to measure those constructs undermines the usefulness of the data collected and limits the types of valid statistical tests and analyses that will be possible.

The SRI survey instrument must be evaluated for validity and bias to ensure that it fair and equitable, and therefore meaningful for summative purposes. Content, criterion, and construct validity should be considered, as well as convergent, discriminant and consequential validity, if applicable. (see Appendix A for a list of definitions of italicized terms)

Assessment and control for biases relating to instructor gender, instructor age, course difficulty, class size, course delivery mode, and non-response rates is also necessary to assure appropriate summative use. In particular, Likert scales are sensitive to central tendency, acquiescence, and social desirability biases.

In terms of the total number of questions, there are no fixed guidelines. That said, literature on survey development best practices often mention that having fewer questions increases response rates. In addition, in research that has discussed methods for increasing response rates, 83% of the universities studied had course evaluations with less than 15 questions. The EPC recommends, in order to encourage higher response rates, reduce fatigue, and minimizing response bias, the total number of questions on the SRI be limited to 15 or fewer when possible.

What Should Students Rate?

The effectiveness of many SRIs is undermined by the inclusion of topics that are more appropriately evaluated by a faculty member’s colleagues or peers. Students are generally not well qualified to rate: appropriateness of course objectives; currency of course content; degree to which readings or other assignments are appropriate; instructor’s subject matter knowledge; validity of assessment techniques; and whether grading standards align with departmental or institutional policy (Benton & Ryalls, 2016; Hoyt & Pallet, 1999).

However, students are the best source of information about their own experiences in a course. In addition, research suggests that students are also one of the best sources of information on the faculty member’s teaching practices.

Formative Feedback

The EPC endorses the view that a primary function of a well-designed student rating instrument should be to provide useful feedback to faculty about their teaching and about student experiences in their courses that can be used for course improvement purposes and not for evaluative purposes.

Summative Evaluation

SRIs for individual courses are designed to give feedback about teaching effectiveness in that course, not teaching effectiveness in general. The general consensus of the literature reviewed for this report is that it is not appropriate to judge teaching effectiveness for personnel and evaluation purposes on the basis
of means reported for an individual instructor using survey items from a single course (Spooren, Brockx, & Mortelmas, 2013).

Summary

The research and literature related to students’ rating of teaching, Benton and Cashin (2012) conclude that, given a well-developed instrument, students are reliable and valid raters of their experiences in the classroom.

SRIs need to be able to function for both formative and summative purposes. The primary purpose should be formative feedback to help faculty improve their teaching practice. Questions on the SRI need to reflect agreed upon definitions of effective teaching at the department level and provide enough specificity so that the feedback is useful. Feedback plus consultation regarding the ratings has been shown to improve subsequent ratings of instruction (Benton & Cashin, 2012). When developing an SRI, the number of questions needs to be limited in order to encourage higher response rates and more valid responses.

5. Analysis of SRI Data

Appropriate analysis of SRI survey data depends on the nature of the survey instrument. Single items that are not aggregated into scales represent an ordinal measurement level, and are therefore limited to less powerful statistical measures. Specifically, the mode or median may be used to indicate central tendency, and the range to indicate dispersion. The use of the mean and standard deviation is not meaningful in regard to ordinal data. For hypotheses testing, the Mann Whitney and the Kruskal Wallis tests are appropriate for use with ordinal data.

If the SRI survey instrument qualifies as a true Likert scale, it is not appropriate to consider individual Likert items as interval data, however aggregate measures of the Likert scale are generally considered to be interval data. As such, more powerful statistical measures may be used, such as the mean and standard deviation for central tendency and dispersion, and the t-test and ANOVA as tests of association. As to individual Likert items, they represent ordinal data for which the mean, standard deviation, the t-test, and ANOVA are not meaningful statistical measures. As ordinal data, however, appropriate statistical measures for individual Likert items include the mode, median, and range.

Best practice suggests that some global questions be included in SRIs related to overall teaching effectiveness (for summative evaluation). Another option is to identify the underlying factor, or scale, structure for the SRI and report aggregated factor/scale scores. Given the formative purpose of an SRI, not all of the questions will be appropriate for evaluation related to promotion or other personnel decisions. A department will need to decide which items or factors will be used for summative review and promotion processes. The EPC recommends keeping the focus of the SRI on formative questions which can be used to support continuous improvement.

If an instructor only teaches one course, then consistent ratings from two different terms of the same course may be sufficient for evaluation. For most instructors, however, ratings from a variety of courses are necessary. (Benton & Cashin, 2012). When reporting such data for review and promotion purposes, it may be appropriate to aggregate across multiple courses that are considered similar (i.e. the courses
have similar norms/characteristics including academic level, class size, subject matter, delivery format, teaching style (lecture based, active learning, etc.) and so forth. It would not be appropriate to report on individual items from particular courses which can fluctuate over time. At least six to eight class ratings should be collected before reliable, summative decisions about effectiveness can be made (Benton & Ryalls, 2016).

It is important to check the response rates across courses and response rate should be reported with any results. Aggregating results across courses may help mitigate against low response rates in any given course. We don’t advocate for a specific response rate because appropriate response rates will vary depending on class or sample size. However, we encourage the adoption of approaches which will help increase response rates (which includes limiting the number of SRI questions), particularly as PSU considers moving to an online SRI delivery method.

In addition to identifying the appropriate items and aggregation schemes for summative reporting, it is important to identify comparison groups. It can be helpful to faculty and evaluators to provide data that contextualize the scores for an individual faculty member. However, we caution against identifying a “bright line” (criterion referenced) that all faculty in a given department are compared against. It will be difficult to defend such decisions without having clear reasons for designating a particular level of performance as the acceptable limit. It is possible, however, to identify an appropriate group of similar courses taking into account level, discipline, class size, quantitative subject matter, etc. (norm referenced). The ideal is to provide results for the course itself (e.g., BIO 101) across all sections (Benton & Ryalls, 2016). However, when a course is taught in both face-to-face and online formats, the different modes of delivery must be taken into account in the formation of comparison groups. A faculty member’s aggregate SRI data should not be presented along with comparison group data unless also reporting the results of appropriate statistical tests (parametric or nonparametric) which indicate whether the faculty member’s ratings are statistically different from the reference group.

Even with well-developed SRIs, comparison groups, and statistical techniques, not all department chairs, committees, and administrators have the background to understand the appropriate interpretations of course evaluations. It is important to present the data clearly and to provide guidance about its use (Boysen et al., 2014).

**Covariates**

In a review of the research on SRIs, Benton and Cashin (2012) noted that well-constructed SRIs are generally reliable and have been found to be valid when compared to other acknowledged measures of teaching (teacher self-evaluation, alumni ratings, and evaluations by trained observers) (Spooren et al., 2013). However, significant but small differences have been found related to:

- **Student:** expected grade, effort/attendance/motivation, gender (interacts with faculty gender)
- **Teacher:** Gender (mixed results), teaching experience, age, personality
- **Course characteristics:** class size, difficulty, discipline, perceived workload, course type (level, major vs. elective, etc.)

This indicates the need to be able to account for these variables either in the SRI system or on the SRI itself.
6. Main Recommendations

a. Define effective teaching
An essential first step in developing any effective tool for the rating of instruction is for departments to define effective teaching. The EPC recognizes the autonomy of academic departments to define effective teaching in their specific contexts. We also recognize that effective teaching will vary across modes of course delivery (e.g., large courses, online courses, seminars, etc.). As discussions are moving forward regarding a campus-wide SRI system, the EPC endorses an approach that will accommodate the different definitions of effective teaching across departments and across modes of delivery. As Benton & Cashin (2012) note, “Effective teaching can be demonstrated in many ways, and no instructor should be expected to demonstrate proficiency in all methods and styles. Moreover, teaching methods may vary, depending on the course content, student characteristics and size of class.” Any SRI system will need the flexibility to accommodate different questions across departments.

b. Follow best practices in instrument development
After departments have defined effective teaching and are ready to develop an SRI instrument, they may consider selecting scales (see definition of a Likert Scale from page 2) from a validated instrument as a starting point and adding any specific departmental questions to those already-developed scales. Focus on what students can report on and what they care about (Spooren et al., 2013). The EPC recommends limiting the total number of SRI questions to 15 or fewer when possible in order to encourage higher response rates. The EPC also recommends that the bulk of the SRI be formative questions designed to encourage continuous improvement. Best practice suggests that some global questions be included in SRIs related to overall teaching effectiveness (for summative evaluation). The EPC recommends that departments decide whether to develop scales for teaching qualities or to use a limited number of single global questions for summative evaluation. Appendix B enumerates the steps for developing effective instruments for collecting student ratings of instruction. In addition to deciding on and piloting the questions and rating scales, it is important to think through the uses of the results at the beginning of the process. Determining how results will be used (faculty only, aggregate over time, program level snapshot) and to whom they will be reported (faculty, department chair, promotion committee) should inform the structure of the rating instrument. It is also important to identify and plan for collecting contextualizing information such as discipline, course level, elective vs. major, etc.

c. Standardize administration and policy
After departments develop appropriate instruments, they also need to determine the best way to administer their SRIs. Administration may vary based on department. However, instructions to students should be consistent across administrations in a given department and should preferably be provided in written form. Based on our review, the EPC recommends giving consideration to encouraging robust responses in both face-to-face and online classes, while maintaining student anonymity.

Departments need to decide on and document their policies regarding course evaluations and their use in faculty evaluation in advance and apply those consistently. Departmental policy should provide guidance for the use of results from courses with low response rates, which courses will be aggregated for evaluation purposes as well as policies around data access. For example, faculty should know whether their evaluations are based on data for all courses, all courses after they’ve been taught at least once, or 10 best courses for all faculty, etc. They need to know in advance if there are courses to be excluded (e.g. small seminars).
d. **Conduct appropriate analysis and present results appropriately**

At minimum, any system for collecting SRI information will need to have the capacity to aggregate data at multiple levels (instructor and program) and over time. It will need to present reports for individual courses and to aggregate across a faculty member's courses over time. It will need to be able to designate comparison groups for courses and faculty and to be able to conduct the appropriate statistical tests to compare instructor results to comparison group results and indicate the areas that are statistically different. Reporting needs to include information for both the instructor and the comparison group, information about statistical tests, as well as the contextualizing data such as class size, level, major, etc. The kinds of analyses described in section 5 (above) as being valid must be adhered to.

e. **Plan for the use of findings**

Departments should decide if evidence will be used formatively or summatively. Formative assessment or evaluation during a course or program provides information useful in improving learning or teaching while the activity is still occurring. Summative assessment or evaluation is conducted at the end of a class or program and is used to make determinations of quality, worth, and/or ability to meet targeted outcomes. We recognize the authority of academic departments to not only construct effective SRIs, but also to decide about how those data will be used by individual faculty and by the departments in review and promotion processes. Academic departments must also meet the expectation that SRIs used for performance evaluation are demonstrably valid and reliable.

f. **Provide guidance for interpretation.**

As noted above, not all faculty and administrators understand the appropriate interpretation of SRI data. In order to be sure that SRIs are used appropriately in faculty review processes, instructions need to be provided about the meaning of statistical significance, how to interpret results and to understand what kind of interpretations are not appropriate (Benton & Cashin, 2012). Administrators and members of personnel committees should use broad categories, rather than try to interpret decimal point differences.

g. **Provide supportive feedback**

Since the primary purpose of SRIs is formative feedback to faculty members, and research supports that getting feedback can help faculty improve their teaching, it is important to identify how faculty will be supported to use their course evaluations for improvement. Departments should discuss and document their own practices for supporting faculty in the formative use of SRI information. OAI, through their teaching consultations, is another avenue for feedback and support. OAI can also provide feedback to departments about how aggregated course evaluations may contribute to departmental assessment efforts.
References


Appendix A
Definitions of Italicized Terms
(All definitions are quoted from Wikipedia, various dates, retrieved February 21, 2018)

Types of Biases

**Acquiescence bias** is a category of response bias in which respondents to a survey have a tendency to agree with all the questions or to indicate a positive connotation. **Acquiescence** is sometimes referred to as "yea-saying" and is the tendency of a respondent to agree with a statement when in doubt.

**Central tendency bias** (sometimes called central tendency error) is a tendency for a rater to place most items in the middle of a rating scale.

In social science research, **social desirability bias** is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of over-reporting "good behavior" or under-reporting "bad", or undesirable behavior.

**Ordinal data** is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories is not known. These data exist on an ordinal scale, one of four levels of measurement described by S. S. Stevens in 1946. The ordinal scale is distinguished from the nominal scale by having ordered categories. It also differs from interval and ratio scales by not having category widths that represent equal increments of the underlying attribute.

Nonparametric Statistical Tests

Mann Whitney Test - In statistics, the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. Unlike the t-test it does not require the assumption of normal distributions. It is nearly as efficient as the t-test on normal distributions. This test can be used to determine whether two independent samples were selected from populations having the same distribution; a similar nonparametric test used on dependent samples is the Wilcoxon signed-rank test.

The Kruskal–Wallis test by ranks, **Kruskal–Wallis H test** (named after William Kruskal and W. Allen Wallis), or one-way ANOVA on ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney U test when there are more than two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA). A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. For analyzing the specific sample pairs for stochastic dominance in post hoc testing, Dunn's test, pairwise Mann-Whitney tests without Bonferroni correction, or the more powerful but less well known Conover–Iman test are appropriate. Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the less stringent
assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

Validity

**Consequential validity** refers to the positive or negative social consequences of a particular test. ... **Consequential validity** was first proposed by Samuel Messick, a psychologist working for the Educational Testing Service.

**Construct validity** is "the degree to which a test measures what it claims, or purports, to be measuring." ... Modern validity theory defines **construct validity** as the overarching concern of **validity research**, subsuming all other types of **validity** evidence.

In psychometrics, **content validity** (also known as logical **validity**) refers to the extent to which a measure represents all facets of a given **construct**. ... A disagreement about a personality trait will prevent the gain of a high **content validity**.

**Convergent validity**, a parameter often used in sociology, **psychology**, and other behavioral sciences, refers to the degree to which two measures of constructs that theoretically should be related, are in fact related. **Convergent validity**, along with **discriminant validity**, is a subtype of **construct validity**.

In psychometrics, **criterion** or concrete **validity** is the extent to which a measure is related to an outcome. **Criterion validity** is often divided into concurrent and predictive **validity**. Concurrent **validity** refers to a comparison between the measure in question and an outcome assessed at the same time.

In psychology, **discriminant validity** or divergent **validity** tests whether concepts or measurements that are not supposed to be related are actually unrelated. Campbell and Fiske (1959) introduced the concept of **discriminant validity** within their discussion on evaluating test **validity**.
Appendix B

Guidelines for best practice in student ratings of instruction (from OAI)

1. Follow best practices in instrument development
2. Standardize administration and policy
3. Plan for use of findings

1. **Follow best practices in instrument development.**
   - Clearly define effective teaching as well as the components
   - Develop questions that measure these components
   - Measure behaviors/practices that are observable
   - Develop response scale
   - Develop two forms: one for formative assessment and one for summative
   - Test questions with the intended audience and use to modify each form
   - Determine how to analyze data appropriately, based on scale and use
   - Determine how results will be reported and used
   - Determine who will provide results and in what form
   - Collect contextualizing information, especially if comparisons are made (discipline, yrs. taught, ..., course elective or required, course level)

2. **Standardize administration and policy**
   - Standardize administration process:
   - Set stage for importance (and use) with students and faculty
   - Use an agreed upon core set of questions
   - Use motivators rather than incentives (consider equity in this online vs face-to-face) to increase student response rates
   - Ensure student anonymity
   - Determine method(s) of delivery (face-to-face or online)
   - Embed in context likely to increase response rate
   - Aim for a response rate of at least 10 students per course or 2/3 of class whichever is higher. If response rates low, combine scores across terms for the same course.
   - Standardize analyses based on data type (e.g., treat ordinal variables appropriately).
   - Provide data interpretation and distribution guidelines
   - Determine audience for results and level of information needed.
   - Determine policy around data storage and access

3. **Plan for use of findings**
   - Develop plans before collecting assessment evidence
   - Develop recommendations for formative use to improve teaching.
   - Develop guidelines for summative use with other evidence to improve decision making.