office of Academic Planning & Assessment
University of Massachusetts Amherst

Martha Stassen
Assistant Provost, Assessment and
Educational Effectiveness
545-5146 or mstassen@acad.umass.edu

Student Response to Instruction (SRTI*) and Performance Appraisal

A Guide for Using SRTI Results to Inform Merit, Promotion, and Tenure

The Student Response to Instruction (SRTI) was developed to provide valid and reliable data on students' experiences in the classroom for two specific purposes: *formative evaluation*, where faculty use data for the improvement of individual teaching practices and *summative evaluation*, where instructors, personnel committees, Departmental Chairs, and Deans use the data as one indicator of teaching performance for merit, promotion, and tenure decisions. This guide is designed to help you effectively incorporate SRTI results into the *summative* evaluation process. In Appendix A, we have included a case study demonstrating how to apply the guidelines for interpreting and understanding SRTI results to an individual instructor's course.

Keep in mind that **student ratings of instruction are only one piece of any evaluation of teaching performance.** Teaching is a complex and multifaceted activity and evaluation specialists recommend considering multiple sources of information to appropriately reflect the various dimensions of overall teaching performance (Cashin, 1999; Centra, 1993; Theall and Franklin, 2001). The Institute for Teaching Excellence and Faculty Development (TEFD) (5-1225, tefd@acad.umass.edu) maintains a collection of materials on the use of additional sources of information to evaluate teaching, including peer review of course materials, classroom observation, and teaching portfolios.

Guidelines for Interpreting Individual Section Reports

At the heart of the SRTI evaluation system is the *Individual Section Report*. For each course section an instructor is teaching in the current semester, we produce a two page report summarizing student responses to each item on the SRTI instrument.

The following guidelines highlight precautions and steps to take in order to ensure the most appropriate and meaningful interpretation of these SRTI results. Applying these guidelines is particularly important for summative evaluations when high-stakes promotion and merit decisions are being made.

1. Focus on results for SRTI Global Items 10 - 12

SRTI items 10 through 12 are "global" items **(Table 1)** and are the items best suited for informing summative evaluations of teaching performance. Research shows that global items are most highly correlated with student achievement and satisfaction and are applicable and comparable in nearly all teaching and learning situations (Centra, 1993). Most sources agree that global items best support decisions related to teaching performance (Abrami, 2001; Arreola, 1995; Centra, 1993).

In contrast, SRTI "diagnostic" items (items 1-9) highlight specific strengths or areas for improvement in teaching performance (i.e., formative evaluation). Though highly correlated with global items, diagnostic items are best used to inform and improve individual practices in specific areas. It is conceivable that an instructor could have a low mean on a diagnostic item and still achieve a very positive overall evaluation.

*SRTI Course Evaluation System is the product of a collaboration between the Office of Academic Planning and Assessment (OAPA), the Institute for Teaching Excellence and Faculty Development (TEFD), and the Faculty Senate Council on Teaching, Learning, and Instructional Technology.

Table 1: SRTI Global Items

10. **Overall, how much do you feel you have learned in this course?**
(5=Much more than most courses, 4=More than most courses, 3=About the same as others, 2=Less than most courses, 1=Much less than most courses)

11. **What is your overall rating of this instructor's teaching?**
(5=Almost always effective, 4=Usually effective, 3=Sometimes effective, 2=Rarely effective, 1=Almost never effective)

12. **What is your overall rating of this course?**
(5=One of the best, 4=Better than average, 3=About average, 2=Worse than average, 1=One of the worst)

2. Review results from multiple courses across multiple semesters

Whenever possible, review student rating results from a variety of courses and from several course sections, spanning at least two semesters (Cashin, 1999; Centra, 1993). Instructors may be more effective in some courses than in others and success in a given course may vary from term to term. Reviewing results from a number of course sections over successive terms will help create a more accurate and meaningful picture of teaching performance over time. For tenure and promotion decisions it is especially important to consider trends over time and to minimize the impact of results for any one particular course.

Because of differences in class size and student class levels, do not combine or average results for several courses taught by an individual instructor. This may mask important differences in the instructor's effectiveness in teaching various types of courses (Theall, 2001).

3. Check sample quality and student characteristics

When evaluating teaching performance, addressing the items in the following checklist will help you determine if the sample of students who responded to the survey is reliable and representative of all students enrolled in the class. Certain course characteristics beyond the instructor's control that may slightly influence ratings are also highlighted and should be taken into consideration when making judgments about instructor effectiveness.

☑ What to note:

For each course section included in the evaluation:

- **What is the return or response rate for the course?** If at least two-thirds of the students enrolled in a course responded to the survey, the results can be considered representative of the entire class (Cashin, 1999; Centra, 1993). If fewer than two-thirds of the students responded to the survey, use caution interpreting the results. A response rate of less than 50 percent indicates the possibility of serious bias and results should not be considered a valid sample of student opinion. (A warning is printed on the reports of sections with response rates below 50 percent.) For courses with small enrollment sizes (5-20 students) use a more stringent response rate criteria.

- **How many students responded to the survey?** If fewer than 10 students evaluated the course, caution should be used in interpreting the results (Cashin, 1999; Centra, 1993). In courses with 10 or more respondents, the effects of a few divergent opinions are limited.

- **How many students are enrolled in the course?** Class size may have a small effect on student ratings (Centra, 1993). Students tend to rank instructors teaching small classes (fewer than 30) higher than instructors teaching larger courses. (Note: The *Individual Section Report* includes means at the department, college, and campus level for courses in the same enrollment category as the section being evaluated.)

- **What are the standard deviations of the three global items?** The standard deviation (SD) is an index of agreement or disagreement among respondents for a particular course. If all respondents agreed exactly (e.g., 100 percent answered “4” on a particular item), the SD would be zero. If the SD for a particular item is greater than 1.20, student responses may be split between high and low ratings, or evenly distributed across response categories (University of Arizona, 2001). In such cases, the mean or average rating does not reflect group consensus on an item and it would be better to examine the item frequencies (the percent of responses in each response category) for a more accurate description of student opinion.

- **Review course characteristics that may have small effects on student ratings.** Did the majority of students in the course take it as a requirement or an elective? Students tend to give slightly higher ratings to courses in their major and electives than to courses taken to fulfill a college or general education requirement. What is the distribution of class levels represented in the class? Higher student ratings are associated with a higher class standing. Lower division students tend to give the lowest ratings and graduate students tend to give the highest ratings.

Criterion and Norm-based Evaluation

There are essentially three approaches to evaluating or drawing conclusions about instructor performance based on SRTI results: a *criterion-based* approach, a *norm-based* approach, or some combination of both. In a criterion-based evaluation, the performance of an individual instructor is compared to a previously determined, fixed standard of excellence (e.g., any mean rating over 4.0 is defined as “Excellent”). Norm-based evaluations are concerned with how the teaching performance of an individual compares to the overall performance of an appropriate group of peers.

Each approach has its advantages and drawbacks. Criterion-based evaluations provide clear standards for teaching excellence independent of the performance of others, but because student ratings of instruction are typically positively skewed (i.e., students tend to rate instructors fairly highly) it may be difficult to derive a set of standards that distinguishes teaching excellence (Abrami, 2001). Norm-based comparisons make it easier to define outstanding performance (e.g. the top ten percent of department faculty are deemed “Outstanding”), but are only appropriate if there is sufficient variation in scores in the comparison group (University of Arizona, 2001).

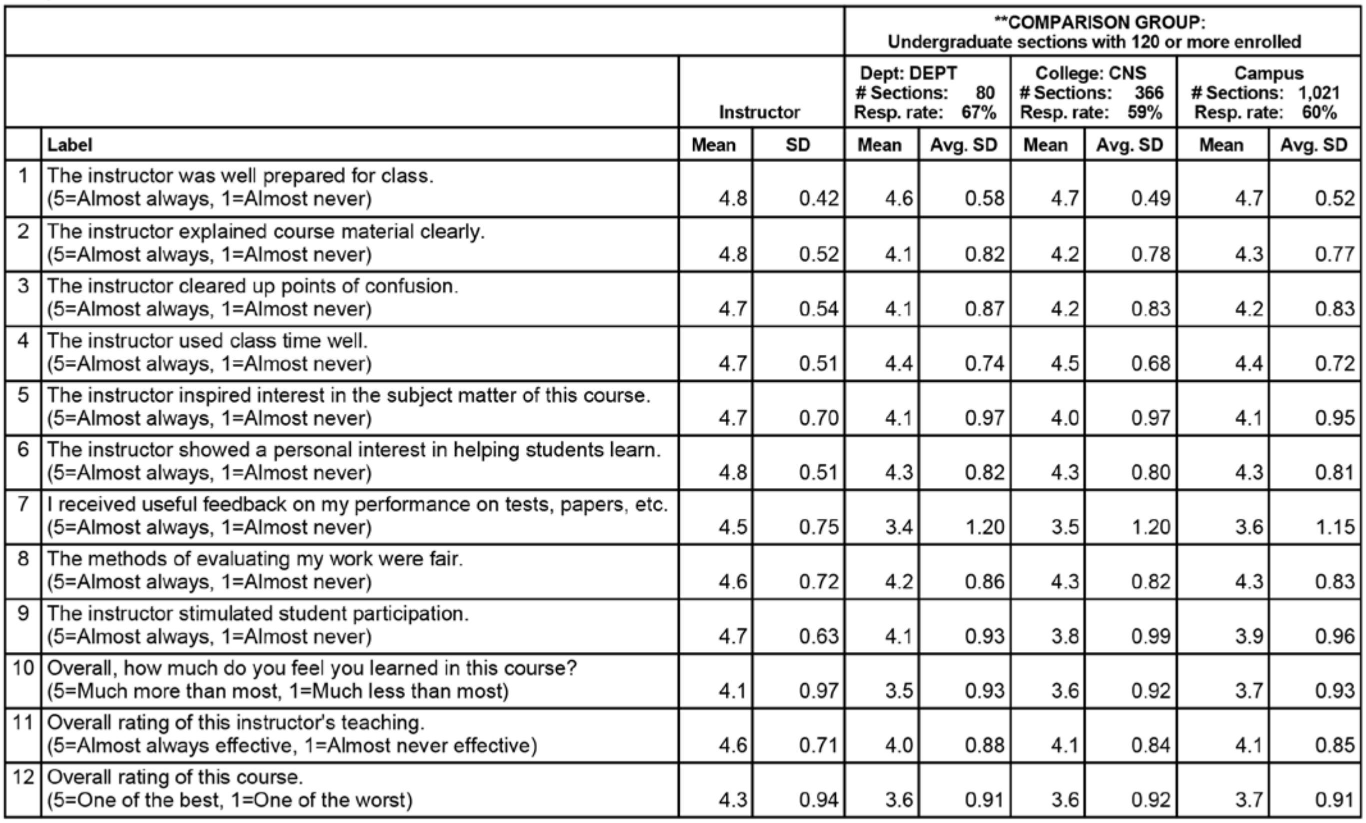
In reality, it is quite common to rely on a combination of both approaches. For example, a criterion-based approach may be developed based on results from historical normative data. You’ll find an example illustrating this approach in the next section.

Tools for Interpreting SRTI Results in Context

1. Individual Section Report Comparison Data

Whatever your approach to summative evaluation, because of the imprecise nature and positive response bias inherent in student ratings of instruction, we include comparison SRTI data on the *Individual Section Report* to help provide a context in which to interpret results for an individual instructor **(Figure 1)**. Item means are provided at the department, school/college, and campus level for all courses in the same class level and enrollment category as the section being evaluated (e.g. undergraduate sections with 120 or more enrolled). The comparison group means are calculated from *combined SRTI data for the three most recent academic years* and are only reported for groups that have ten or more sections. The comparison group data do not include courses that are fewer than 2 credits, noncredit labs or discussions, independent study, practicum, or dissertation sections.

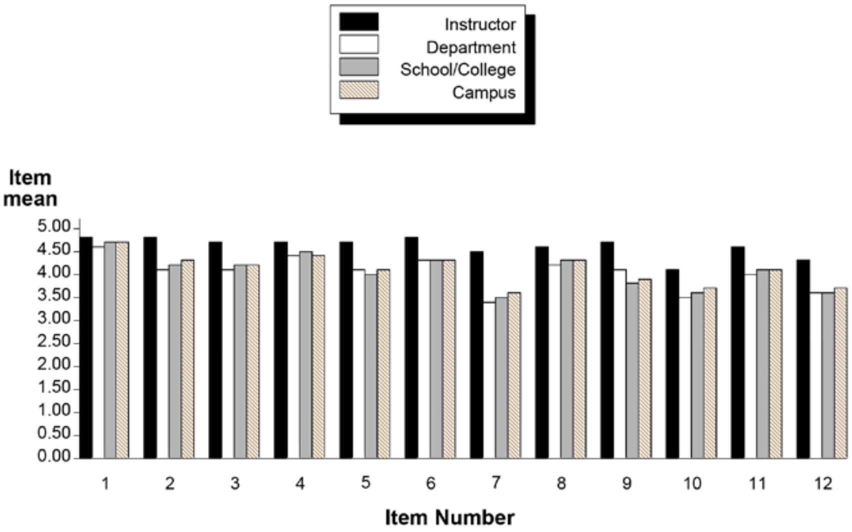
Figure 1: Item Means – SRTI Individual Section Report Page Two (Excerpt)



				**COMPARISON GROUP: Undergraduate sections with 120 or more enrolled					
		Instructor		Dept: DEPT # Sections: 80 Resp. rate: 67%		College: CNS # Sections: 366 Resp. rate: 59%		Campus # Sections: 1,021 Resp. rate: 60%	
	Label	Mean	SD	Mean	Avg. SD	Mean	Avg. SD	Mean	Avg. SD
1	The instructor was well prepared for class. (5=Almost always, 1=Almost never)	4.8	0.42	4.6	0.58	4.7	0.49	4.7	0.52
2	The instructor explained course material clearly. (5=Almost always, 1=Almost never)	4.8	0.52	4.1	0.82	4.2	0.78	4.3	0.77
3	The instructor cleared up points of confusion. (5=Almost always, 1=Almost never)	4.7	0.54	4.1	0.87	4.2	0.83	4.2	0.83
4	The instructor used class time well. (5=Almost always, 1=Almost never)	4.7	0.51	4.4	0.74	4.5	0.68	4.4	0.72
5	The instructor inspired interest in the subject matter of this course. (5=Almost always, 1=Almost never)	4.7	0.70	4.1	0.97	4.0	0.97	4.1	0.95
6	The instructor showed a personal interest in helping students learn. (5=Almost always, 1=Almost never)	4.8	0.51	4.3	0.82	4.3	0.80	4.3	0.81
7	I received useful feedback on my performance on tests, papers, etc. (5=Almost always, 1=Almost never)	4.5	0.75	3.4	1.20	3.5	1.20	3.6	1.15
8	The methods of evaluating my work were fair. (5=Almost always, 1=Almost never)	4.6	0.72	4.2	0.86	4.3	0.82	4.3	0.83
9	The instructor stimulated student participation. (5=Almost always, 1=Almost never)	4.7	0.63	4.1	0.93	3.8	0.99	3.9	0.96
10	Overall, how much do you feel you learned in this course? (5=Much more than most, 1=Much less than most)	4.1	0.97	3.5	0.93	3.6	0.92	3.7	0.93
11	Overall rating of this instructor's teaching. (5=Almost always effective, 1=Almost never effective)	4.6	0.71	4.0	0.88	4.1	0.84	4.1	0.85
12	Overall rating of this course. (5=One of the best, 1=One of the worst)	4.3	0.94	3.6	0.91	3.6	0.92	3.7	0.91

A bar chart is displayed at the bottom of page two of the *Individual Section Report* to provide a visual representation of the instructor and comparison group means on each item. **(Figure 2).**

Figure 2: Item Means – SRTI Individual Section Report Page Two (Excerpt)



☑ What to note

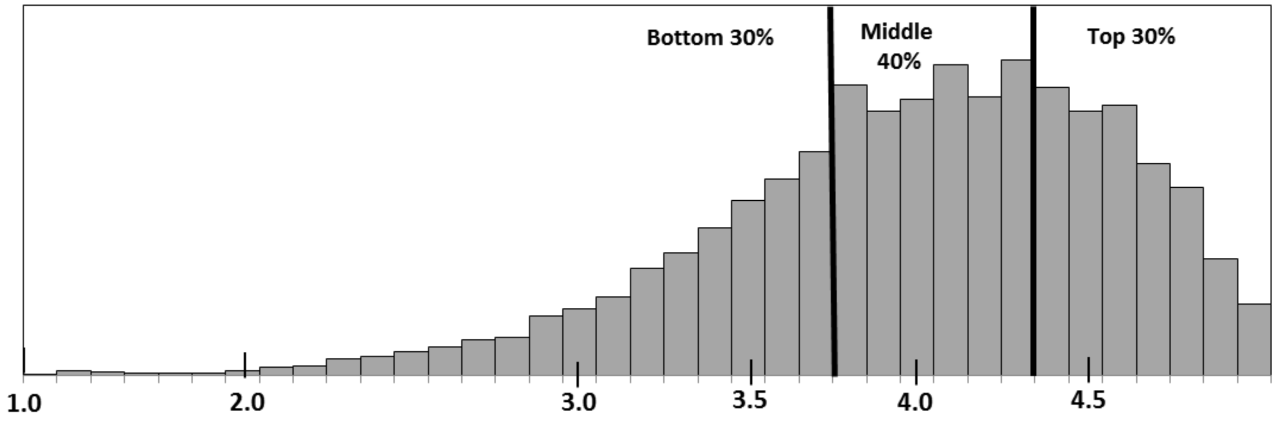
- Do not treat norm group results as an absolute standard or line of demarcation between “passing” and “failing” instructor performance.

- Also avoid using trivial differences in mean scores to rank or compare the teaching performance of individual instructors.

- Whenever you compare results for an individual instructor to the group results presented in these reports, take care not to over-interpret small differences in mean scores. In general, a difference smaller than .4 does not represent a real difference in teaching performance (University of Arizona, 2001).

Student ratings of instruction are best at distinguishing the *extremes* - that is, instructors whose ratings are well below or well above average. They lack the precision for ranking or making fine distinctions among the majority of instructors whose ratings fall somewhere in the middle. **Figure 3** shows the distribution of means for all undergraduate sections on global item 12, "*What is your overall rating for this course.*" Notice the middle 40 percent of sections have means somewhere between 3.8 and 4.3. In this case, a mean of 3.9 for one instructor is not meaningfully worse than a rating of 4.3 for another instructor. Figure 3 also demonstrates the positive response bias typical in student ratings. Though the midpoint of the 5-point response scale is 3.0, "About average", this should not be interpreted as an "average" rating; over 80% of courses had a mean rating of 3.5 or higher, for an average rating of 4.0, a full point above the item midpoint.

Figure 3: Distribution of Course Means for Item 12
All Undergraduate Courses: Academic Years 2013 - 2015



2. Summary Statistics for SRTI Global Items Report

Appendix B of this document contains the *Summary Statistics for SRTI Global Items* report, which shows mean and percentile distributions for each global item for the entire campus and for each school/college. Percentiles are calculated for three academic years (2013, 2014, and 2015) worth of SRTI data. In addition, means and percentiles are given for the four categories of student enrollment and for undergraduate and graduate courses.

One way to avoid over-emphasizing small differences in mean scores is to assign an instructor's SRTI means to one of 3 to 5 previously determined categories of performance (Arreola, 1995; Centra, 1993). Arreola (1995) suggests using comparison group percentile scores to determine these categories. In **Figure 4,** we have modified an excerpt from the *Summary Statistics for SRTI Global Items* report to define 5 categories of performance: 'Much Lower', 'Lower', 'Similar', 'Higher', and 'Much Higher'.

An item mean that falls below the 10th percentile is categorized as 'Much Lower.' A mean that falls somewhere between the 10th and 30th percentiles is categorized as 'Lower', and so on. So for example, an instructor who received a mean rating of 3.9 on item 11 falls between the 30th and 40th percentiles and would be categorized as 'Similar'. In other words, the instructor's mean on item 11 falls in the middle 40% of means for instructors who taught undergraduate sections with 60 to 119 students enrolled during academic years 2013-15.

* A *percentile* represents the point in a distribution at or below which a given percentage of responses fall.

Figure 4: Summary Statistics - Undergraduate Sections, Academic Years 2013-15

	ENROLLMENT OF 60 - 119							No. of course sections:		1111				
				Percentiles										
				Much Lower	Lower		Similar				Higher		Much Higher	
	Item	Mean	SD	Min	10%	20%	30%	40%	Median	60%	70%	80%	90%	Max
Q10	Overall, how much do you feel you have learned in this course?	3.69	0.56	1.30	2.92	3.26	3.46	3.59	3.75	3.88	4.02	4.19	4.37	4.84
Q11	What is your overall rating of this instructor's teaching?	4.08	0.60	1.69	3.23	3.62	3.86	4.03	4.21	4.35	4.46	4.58	4.73	5.00
Q12	What is your overall rating of this course?	3.70	0.62	1.30	2.87	3.21	3.43	3.61	3.77	3.91	4.07	4.23	4.42	5.00

Summary

Student perceptions of teaching effectiveness are a valuable component of any evaluation of teaching performance. However, while students can appropriately evaluate aspects of teaching that reflect student experience with an instructor (e.g., student-instructor relationships, instructor ability to communicate clearly, fairness of grading), they are not the best judges of other components of teaching, such as instructor subject matter expertise (e.g., knowledge in major field, course syllabus and reading list, selection of course objectives and materials). It is essential that personnel committees use SRTI results in conjuction with other sources of information to evaluate teaching perfomance. The Center for Teaching and Faculty Development maintains a collection of materials on incorporating additional sources of information, such as peer review of course materials, classroom observation, and teaching portfolios, for individual and departmental use.

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research, No. 109. San Francisco: Jossey-Bass.

Arreola, R.A. (1995). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker Publishing Company.

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

Cashin, W.E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin & Associates, *Changing practices in evaluating teaching*. Bolton, MA: Anker Publishing Company, Inc.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research, No. 109. San Francisco: Jossey-Bass.

Theall, M. (2001). Can we put precision into practice? Commentary and thoughts engendered by Abrami's "Improving judgments about teaching effectiveness using teacher rating forms". In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research, No. 109. San Francisco: Jossey-Bass.

University of Arizona, Assessment and Enrollment Research (2001). Guide to student ratings at the University of Arizona.