Feature Article

How Students Rate, Part 2: Pilot Assessment of Relevance of Student Reasoning in Course Ratings

Christine Winskowski
Morioka Junior College, Iwate Prefectural University

Susan Duggan
Iwate Prefectural University

This paper follows Part 1 (Winskowski, 2010), in which 10 students were interviewed while completing a course evaluation form for understanding of items and reasons for rating. While most responses seemed reasonable and predictable, a significant minority contained problematic interpretation, idiosyncratic reasoning, etc. To examine this more directly in Part 2, 8 college-level instructors assessed transcribed student responses from Part 1 as Relevant to the item, Irrelevant, or Unclear. Overall, instructors found 72.6% of responses Relevant, with 27.4% as Irrelevant or Unclear. The 10 students' response sets ranged from 56.3% to 82.3% in the mean Relevance assessments received from the 8 instructors, and the 12 items ranged from 42.5% to 97.5% in the mean Relevance assessments received. Instructors ranged from 54.2% to 85.8%.in the Relevance assessments they gave. The percentage of agreement between instructor pairs ranged from 48.3% to 75.8%, with a mean agreement of 63.9%. Variable levels of relevance attributed to items, to students' responses, and variable latitude in the instructors' assessments casts doubt on the validity of results from a conventional

Winskowski, C., & Duggan, S. (2011). How students rate, Part 2: Pilot assessment of relevance of student reasoning in course ratings. *OnCUE Journal, 4*(3), 217-242.

student ratings instrument.

本稿はPart1 (Winskowski, 2010)の続編である。前編では、10名の学生による講座評価調書への記入を実施すると同時に、評価項目の理解度及びその評価に至った理由を問う面接を行った。大部分の回答は合理的かつ予測可能なものであったが、有意な少数回答に問題のある解釈や特異な推論等が見られた。この現象を更に直接的に考証するため、Part 2では8名の大学教師により、文書化された前編の学生面接回答のそれぞれについて、評価項目との関連性が妥当(relevant)、無関係(irrelevant)、不明(unclear)の評価を行った。全体としては、72.6%の回答がRelevant、27.4%がIrrelevant またはUnclearと評価された。10セットの学生回答は、8名の評価者からの受けた妥当評価の平均値56.3% から 82.3%を示した。また、12アイテムそれぞれの妥当評価の平均値は42.5% から 97.5%であった。評価を行った教師間では、Relevantの評価に54.2% から85.8%の開きが見られた。評価者の組み合わせによる評価の一致率は48.3% から75.8%で、平均は63.9%であった。評価項目や学生の回答によって妥当性が大きく変動すること、また評価を行った教師間に見られる評価結果のばらつきは、慣習的に行われている学生評価手法の結果の有意性について疑問を投げかけるものである。

Conventional Student Ratings Instruments (SRIs) are widely used to evaluate university courses and instructor effectiveness in the western hemisphere, increasingly in Asia and other parts of the world (Marsh, 2007), and now in a growing number of colleges and universities in Japan. A very large research literature on the characteristics of SRIs (also called Student Evaluations of Teachers or SETs) has emerged in the last several decades with a large portion focused on quantitative measures of validity, e.g. linking SRIs with grades and grade expectations, exam scores, sources of bias, instructor personality traits, course characteristics (workload, course size, level, discipline), student characteristics (reason for taking a course, prior interest). While SRIs are widely held to be moderately valid (Abrami, d'Apollonia & Cohen, 1990; Cashin, 1995; McKeachie, 1997), a smaller but consistent countercurrent has emphasized the flawed nature of these instruments (e.g., Birnbaum, 1999; Clayson & Sheffet, 2006; Johnson, 2003; Orsini, 1986). Sufficient concern about SRIs and advice against their sole use

in faculty evaluation has been voiced through the years, so that when Selden notes in a 1998 survey that 88.1% of institutions in the U.S. use SRIs for evaluation of faculty effectiveness, he also describes increases in the simultaneous use of alternative forms of faculty evaluation by chairs, deans, and colleagues; classroom observation; peer review; materials portfolios; and other qualitative measures (Selden, 1999). Thus, the preponderance of research and recent practice in evaluation of instructor effectiveness reflects continued use of quantitative analysis on the statistical features of SRIs, with the growing addition of qualitative alternatives in documenting instructor effectiveness.

Absent in the quantitative focus, and bypassed in the qualitative alternative, is the documentation of patterns and features shown in students' thinking and reasoning the processes while filling out course evaluation forms. It is the study of these patterns and features which provides direct evidence of the validity of an evaluation instrument. That is, it reveals students' discoursal understanding of items and the discoursal premises of their responses, thereby telling us what we can and cannot expect to understand from the results of our students' course and teacher ratings.

This paper describes the second phase of a pilot study of students' thinking processes as they complete standardized, 12-item college course-evaluation forms. In Part 1 (Winskowski, 2010), as students filled out a course-evaluation form, they were interviewed and asked to explain their understanding of each evaluation item (e.g. "I have had a strong interest in this course from the beginning.") and the reasons for selecting their rating, ranging from "1 = not appropriate" to "6 = appropriate." While the majority of the students' responses seemed expected and even predictable, a significant minority appeared to include unwarranted assumptions, problematic interpretation of the items, idiosyncratic reasoning, and other difficulties.

The distinction between student statements that appeared responsive to the items and those that did not, and what might be revealed in that

distinction, merited further investigation. In Part 2 of this study, a small group of instructors including the authors sorted the students' item responses according to whether they appeared Relevant to the item, Irrelevant, or Unclear (as to relevance) to see how these distinctions might be observed by others and to explore patterns in the data based on these observations.

This work adds to a small body of recent research which looks directly at students' thinking and reasoning processes as they complete ratings forms. Indeed, several of the results in the first part of this study echo findings from these studies, showing that the appearance of objectivity from numerical student ratings (i.e. scores, instrument means, class means, etc.) masks the portion of students' responses which differ from an assumed norm of reasoning and rating selection. Like the data from a comprehensive study by Benz and Blatt (1996) which asked students to write their reasons for choosing an item rating, Part 1 showed that students sometimes rated the subject matter (rather than teaching effectiveness), and that some items yield more unified themes in reasoning than others. Like the results reported in both Benz and Blatt (1996) and in Billings-Gagliardi, Barrett, and Mazor (2004) who conducted "think-aloud" interviews as students completed a course evaluation, as well as Burden (2008b) who conducted teacher-interviews on the subject, the data also showed that students attributed some ratings to their own ability (rather than teaching effectiveness) and offered explanations for their ratings which instructors may regard as idiosyncratic or unexpected. Like Kolitch and Dean's (1998) study which examined students' explanations of global item ratings, Part 1 showed that students employed multiple interpretations of a global item. The significance of students' understanding and interpretation of evaluation items cannot be underestimated. Benz and Blatt (1996) point out that "…understanding the students' interpretations of the items used is crucial. This concept gets to the heart of validity…" (p. 429). Like Benz and Blatt, we hope to contribute to an understanding

of the construct validity of SRIs in general, as well as to add to the literature on students' course evaluation in Japan's universities.

Method

In the first phase of the study (Winskowski, 2010), 10 second-year students were interviewed as they completed a conventional, college-mandated 12-item rating form to evaluate a second-year, foreign language course emphasizing speaking and listening of English. They were asked (in their native Japanese): 1) What does this item mean? and 2) Why did you give the item that rating? As is indicated above, the dichotomy between student answers that seemed item-responsive and item non-responsive provided the impetus for Part 2 of this study. More specifically, most students' responses to evaluation items seemed sensible, reasonable, and even predictable. However, there were a substantial number of student interview responses that revealed premises that were unwarranted, e.g., that some condition like "looking at and knowing the syllabus" was fulfilled (i.e. for an item asking whether the content of the course corresponded to the syllabus), when it was not, and that all students interpreted the items as instructors and presumably item-writers would, when some did not. It might be assumed that students understood the terminology of the items, whereas some students clearly had difficulty in understanding, and that students determined ratings in a manner consistent with one another, when they actually chose different criteria. Further, we found that students sometimes chose a lower rating based not on instructor-effectiveness, but rather on their own effectiveness as students. Intrigued by how aspects of our data differed from what might be expected (and what colleagues and administrators might expect), we decided to look more directly at the relevancy of responses, that is, whether a response seemed to appropriately and meaningfully address the item as we understood it.

First, the authors attempted to see informally if we could agree on whether student responses were relevant, irrelevant, or not clearly either. We independently assessed each student response as relevant (O) to the item, irrelevant (X), or not clearly either (?), finding agreement on 76.6%[1] of student responses: 76 out of 120 or 63.3% appeared relevant to both of us, 16 or 13.3% appeared irrelevant, and none appeared indeterminate. This measure of agreement was less than anticipated, with an unweighted Cohen's kappa value of .40[2] suggesting low to moderate agreement at best.

A re-assessment of student responses was made, this time first making item paraphrases to give us a common set of references. An increase in agreement was anticipated, but in fact, it decreased slightly, resulting in agreement on 73.3% of the student responses (70 out of 120 or 58.3% appeared relevant to both of us, 17 or 14.2% appeared irrelevant, and one 1 or .01% appeared indeterminate). There was no appreciable change in the Cohen's kappa value.

If a significant portion of student responses (26.7% and 23.4% for the first and second attempts) could not be agreed upon by us as responding relevantly to the items, it certainly seemed to raise doubt about the validity of the SRI that was used and the responses they prompted. To examine the question of how our assessments of student responses would compare to colleagues' assessments, in this study we extended our original attempt by asking a small group of six colleagues to repeat the assessment process in order to see what percentage of the responses they would find relevant, irrelevant, or not clearly either.

Participants

A total of 8 assessors participated, comprising 4 university or college English instructors (including the authors), 1 English instructor of a private language school, 1 college social science faculty member, and 2 individuals who worked as translators and taught English or Japanese

part-time in universities and privately. Two assessors (the authors) were English native speakers, and the remainder Japanese native speakers with advanced or near-native ability in English. As this was a pilot effort, we approached colleagues whom we knew, who were familiar with the student ratings process, and who would be interested in the study and sympathetic with its aims.

Materials

Each assessing instructor was provided with transcripts of the interview responses grouped by item (i.e. all responses to Item 1, all responses to Item 2, etc.), either in the original Japanese or English translations, as they preferred. All assessors but one Japanese individual chose transcripts in their native language. Assessors were also supplied with a grid, with student numbers (1-10) on the horizontal axis, and SRI item numbers (1-12) on the vertical axis (see Appendix). These grids were to be used to record the assessments of student responses as relevant, irrelevant, or not clearly either.

Procedures

The authors' original assessments were treated as part of the data collection since it fit the conditions of other instructors' assessments, namely no discussion of item definition prior to assessing. No special definitions of the term "relevance" or other terms were offered to the participating instructors since college and university instructors are not ordinarily provided glosses in the course ratings process or on receipt of results. Following the original procedure, the instructors were asked to indicate, according to their own judgment, whether a response seemed Relevant to the item in the usual sense (marked with "O"), Irrelevant ("X"), or Unclear (not clearly Relevant or Irrelevant) ("?").

To illustrate, examples of individual student responses which were most frequently categorized as Relevant, Irrelevant, and Unclear are shown below:

a. Example response assessed as Relevant (O) by all 8 instructors

Item 12. Considering the course as a whole, I feel satisfied with it.

Student 1: Rating - 6

Yes. We used a variety of materials, studied a lot of English and it was fun...and it was good to be able to talk about ourselves to one another in English.

b. Example response assessed as Irrelevant (X) by 7 of 8 instructors

Item 2. The content of the lessons corresponded to the syllabus.

Student 4: Rating – 4

I haven't read the syllabus much so I don't really know but if I compare it to the First Year, we are doing the same sort of things so it might probably correspond to the syllabus, I think.

Interviewer: It might probably correspond so you put "4"?

Student 4: Yes.

c. Example response assessed as Unclear (?) by 3 of 8 Instructors

Item 4. The use of teaching materials (blackboard, audio-visual aids, textbook, handouts etc) was appropriate.

Student 7: Rating - 5

This is "5".

Interviewer: What do you think "The use of teaching materials... was appropriate" means?

Student 7: Did you use the textbook properly as the course progressed?...The textbook we are using now is really easy to understand so we followed it through the course, "5 - This statement is mainly appropriate."

Interviewer: Alright, but it's not "6"?

Student 7: No, it isn't. Often, the textbook...we didn't only use the textbook every lesson, so "This statement is mainly appropriate."

Interviewer: Alright. If you used the textbook more, would you choose a higher number?

Student 7: Yes.

The instructors' assessments of students' item-responses were entered into a table, exemplified in Table 1. Here Student 1's assessment data (O, X, and ?) for Items 1 to 12 are shown for each instructor, A – H. The columns on the right show the percentage of assessments comprising O's, X's, and ?'s. For example, for Item 1, 5 out of 8 instructors assessed Student 1's response as Relevant (O), a percentage of 62.5%. Two instructors assessed the response as Irrelevant (X), a percentage of 25%, and 1 as Unclear (?), 12.5%.

Thus, each student response for the 12 items (10 students X 12 items = 120 responses) was similarly assessed for relevance by 8 assessors (120 responses X 8 assessments) for a total of 960 assessments. Then, for each student, a mean percentage of O's, X's, and ?'s was derived for all Items 1-12. As can be seen above, Student 1 had mean O assessment percentage of 82.3% for all 12 items (i.e. seemed Relevant to the assessors), a mean X assessment percentage of 14.6% (seemed Irrelevant), and a mean ? assessment of 3.1% (seemed Unclear). Identifying mean assessment percentages for each student offers a way that students might be compared with one another, i.e. to see if a given student's responses seemed to assessors as relevant as other students' responses.

Next, for each *item*, a similar mean percentage of O assessments, X assessments, and ? assessments was derived for responses from Students 1-10. Identifying mean assessment percentages for each item offers a way of comparing items to see if some items might lend themselves to greater ease of student understanding and interpretation. Items that are readily understood by students may be expected to result in more relevant responses than items that are not.

Finally, the distribution of *instructors*' assessments was examined. Instructors were compared for the number and percentage of O's, X's, and ? assessments they made. This allowed us to see whether instructors employed greater or lesser latitude in their assessment of student response relevance. Additionally, the degree to which

Table 1. Student 1's Assessment Data from Instructors A-H with Total Percentages of Relevant (O), Irrelevant (X), and Unclear (?) Assessments

Student 1	Instructors								Assessment Percentage		
	A	B	C	D	E	F	G	H	O%	X%	?%
Item 1- I have had a strong interest in this course from the beginning.	X	X	O	O	O	?	O	O	62.5	25.0	12.5
Item 2 - The content of the lessons corresponded to the syllabus.	?	O	O	O	O	O	?	O	75.0	0.0	25.0
Item 3 - The way of teaching focused on the important points.	X	X	O	O	X	O	O	O	62.5	37.5	0.0
Item 4 - Use of teaching materials was appropriate.	O	O	O	O	O	O	O	O	100.0	0.0	0.0
Item 5 - The quantity of the lesson content was appropriate.	X	O	O	O	X	X	O	O	62.5	37.5	0.0
Item 6 - The teacher's explanations were easy to understand.	O	O	O	O	O	O	O	O	100.0	0.0	0.0
Item 7 - The lesson content was easy to understand.	O	O	O	O	X	O	O	O	87.5	12.5	0.0
Item 8 - The teacher provided opportunities for comments and questions; responded appropriately.	O	O	O	O	O	O	O	O	100.0	0.0	0.0
Item 9 - I was able to sense the teacher's enthusiasm.	X	X	X	O	X	O	O	O	50.0	50.0	0.0
Item 10 - I felt intellectually stimulated by this course.	O	X	O	O	O	O	O	O	87.5	12.5	0.0
Item 11 - I acquired a lot from this course.	O	O	O	O	O	O	O	O	100.0	0.0	0.0
Item 12 - Considering the course as a whole, I feel satisfied with it.	O	O	O	O	O	O	O	O	100.0	0.0	0.0
Students' Means									82.3	14.6	3.1

instructors' assessments agreed with one another was examined by finding the percentage of overlap between each possible pair. More sophisticated analyses of the degree of overlap between pairs or among all instructors are available, e.g. Fleiss' kappa, but the size and pilot nature of this study and the small numbers of participants merited a simpler and more direct approach.

These procedures generated the core of our results. The next section reports on the following analyses:

1. Total distribution of all relevant (O), irrelevant (X), and unclear (?) assessments of student responses.
2. Comparison of students on their mean percentages (for all items) of O, X, and ? assessments.
3. Comparison of items of their mean percentages (for all students) of O, X, and ? assessments.
4. Comparison of instructors on the number and percentage of O, X, and ? assessments across students responses to items.
5. Percentage of assessment agreement between each possible pair of instructors.

Results

The results are presented in order of the analyses, starting with the distribution of instructors' assessments across all data. They continue with comparing results for students, items, and for the instructors themselves.

Total distribution of Relevant (O), Irrelevant (X), and Unclear (?) assessments of student responses by instructors

Each of 10 students gave 12 item-responses, making 120 responses altogether. Each of these responses was in turn assessed by 8 instructors, making a total 960 assessments. Of these 960 assessments, a total of 697 (72.6%) were assessed as Relevant (O), 177 (18.4%) as Irrelevant (X), and 86 (9%) as Unclear (?).

Table 2. For Students 1-10, Mean Percentages for O, X, and ? Responses to All Items, Rank-ordered by Relevance (O)

	% of Assessor Means		
	O	X	?
Student 1	82.3	14.6	3.1
Student 10	79.2	11.5	9.4
Student 3	77.1	11.5	11.5
Student 8	76.0	13.5	10.4
Student 6	71.9	19.8	8.3
Student 7	70.8	19.8	9.4
Student 5	70.8	21.9	7.3
Student 7	70.8	19.8	9.4
Student 4	69.8	20.8	9.4
Student 2	65.6	21.9	12.5
Student 9	56.3	30.2	13.5
Mean of means	71.9	18.7	9.5

Comparison of Students 1-10 on mean percentages (for all items) of O, X, and ? assessments

Calculating the average percentage of instructor O, X, and ? assessments for each student offers a way to compare students' relevance "profiles." Table 2 presents students' mean percentages for O, X, and ? assessments, rank ordered by students with the highest to the lowest Relevance (O) means.

Mean percentages of Relevance (O) assessments ranged from Student 1's high of 82.3% to Student 9's low of 56.3%, a range of 26 percentage points. It can also be seen that generally students with higher mean Relevance assessments tended to have lower Irrelevance assessments, and those with lower mean Relevance assessment had higher mean Irrelevance assessments. That is perhaps unsurprising (though not inevitable, since the third possibility is an assessment of Unclear). The pattern of Unclear (?) assessments seems less definite. As might be expected, the means of students' Relevance (O), Irrelevance (X), and Unclear (?) assessments are 71.9%, 18.7%, and 9.5%, closely

Table 3. For Items 1-12, Mean Percentages for O, X, and ? Responses for All Students, Rank-ordered by Relevance (O)

	% of Assessor Means		
Items	O	X	?
Item 6 - The teacher's explanations were easy to understand.	97.5	1.3	1.3
Item 11 - I acquired a lot from this course.	93.8	2.5	3.8
Item 8 - The teacher provided opportunities for comments and questions, responded appropriately.	86.3	8.8	5.0
Item 12 - Considering the course as a whole, I feel satisfied with it.	83.8	12.5	3.8
Item 9 - I was able to sense the teacher's enthusiasm.	82.9	12.5	5.0
Item 5 - The quantity of the lesson content was appropriate.	72.5	17.5	10.0
Item 1- I have had a strong interest in this course from the beginning.	70.0	16.3	13.8
Item 7 - The lesson content was easy to understand.	68.8	18.8	12.5
Item 4 - Use of teaching materials was appropriate.	57.5	26.3	16.3
Item 10 - I felt intellectually stimulated by this course.	56.3	28.8	15.0
Item 3 - The way of teaching focused on the important points.	52.5	32.5	15.0
Item 2 - The content of the lessons corresponded to the syllabus.	42.5	45.0	12.5
Mean of means	72.0	18.6	9.5

Note: Note that these item means will not sum to 100, since they are each calculated from assessments of the 10 students' item-responses, which are independent of each other.

akin to the overall percents noted above (Result 1) of all 960 Relevance assessments (72.6%, 18.4%, and 9.0% respectively). The most notable finding here is that the range of Relevance and Irrelevance means for students makes it clear that, to the instructors, some students may seem more capable than others at providing evaluation ratings through appropriate and expected, that is, relevant, reasoning.

Comparison of Items 1-10 on mean percentages (for all students) of O, X, and ? assessments

Next, we examine the average percentages of instructors' (O), (X), and (?) assessments for Items 1-12, which allows a comparison of items' relevance "profiles." Table 2 shows the item means, ranked ordered by items from highest to lowest Relevance (O) mean.

Student responses for Item 6 (The teacher's explanations were easy to understand) have a remarkably high mean Relevance assessment of 97.5%. This suggests that this item is sufficiently clear and uncomplicated to students' interpretation that their responses are almost unanimously considered direct and relevant. Student responses for Item 11 (I acquired a lot from this course) also received a high mean assessment of Relevance at 93.8%. Student responses to Items 8 (The teacher provided opportunities for comments and questions and responded appropriately), Item 12 (Considering the course as a whole, I feel satisfied with it), and Item 9 (I was able to sense the teacher's enthusiasm) received mean Relevance assessments in the 80's. In contrast, the lowest mean Relevance assessment was for Item 2 (The content of lessons corresponds to the syllabus) at 42.5%. As the item means for Os decrease, the means for Xs increase (with the exception of Item 1), as we might expect. Generally, the trend of ?s is to increase as well. If the item means for Relevance, Irrelevance, and Unclear assessments are examined for their mean, i.e. the mean of item means, we find they are again similar (72.0%, 18.6%, and 9.5%) to the overall percentages (72.6%, 18.4%, and 9.0%, respectively).

The striking finding here is the dramatic range of item Relevance (O) percentages, from Item 6 at 97.5% to Item 2 at 42.5%, a range of 55 percentage points, and the range of item Irrelevance (X) percentages, from Item 6 at 1.3% to Item 2 at 45.0%, a range of 43.7 percentage points. This certainly suggests that, from the students' perspectives, some items, such as Items 6 and 11, may be easily interpretable and easy to respond to in a clear, relevant way. Others, such as Items 4, 10, 3, and 2 may be harder for students to interpret, and correspondingly more difficult to respond to in a way that is seen as expected, normative, and comprehensible.

Comparison of instructors on their overall number and percentage of O, X, and ? assessments

Next, the degree to which assessments differ among Instructors A through H is addressed. The overall number and percentage of Relevance (O), Irrelevance (X), and Unclear (?) assessments (out of 120 assessments) for each instructor are compared. Table 4 shows the data in rank order of instructors with highest to lowest Relevance (O) percentage.

Once again, a considerable range in instructors' assessments of all three kinds can be seen. Among Relevance assessments, instructor F considered nearly 86% of student responses to be relevant, but Instructor E felt that only about 54% of responses were relevant, a range of nearly 32 percentage points. Assessment of Irrelevant responses ranged from a low of 1.0% (Instructor G) to a high of 39.2% (Instructor E), a difference of 38 percentage points. With two exceptions (Instructors H and G), as the percentage of Relevance assessments decreased, the percentages of Irrelevance assessments increased, as might be expected. Responses that were seen as Unclear ranged from 0 (Instructor C) to 31.7% (Instructor G), but seem to have no clear trend. However, once again, the mean of instructor means for assessments of Relevance, Irrelevance, and Unclarity (72.4%, 18.0%. and 9.6%) falls close to the overall percentages (72.6%, 18.4%, and 9.0% respectively).

Table 4. Numbers (and Percentages) of Assessments of Student Responses (O, X, and ?) by Instructors, Rank-ordered by Relevance (O)

	Responses Assessed as		
Instructor	Relevant - O	Irrelevant - X	Unclear - ?
F	103 (85.8%)	2 (1.7%)	15 (12.5%)
D	97 (80.8%)	21 (17.5%)	2 (1.7%)
A	94 (78.3%)	23 (19.2%)	3 (2.5%)
H	90 (75.0%)	6 (5.0%)	24 (20.0%)
B	89 (74.2%)	29 (24.2%)	2 (1.7%)
C	87 (72.5%)	33 (27.5%)	0.0
G	70 (58.3%)	12 (1.0%)	38 (31.7%)
E	65 (54.2%)	47 (39.2%)	8 (6.7%)
Mean	86.9 (72.4%)	21.6 (18.0%)	11 .5 (9.6%)

Percentage of assessment agreement between each pair of instructors

To determine the degree of agreement of instructors' assessments, each pair of instructors' assessments was examined for a simple percentage of agreement. Table 5 shows pair-wise comparison of instructors' percentage of agreement. Each percentage shows the combined total agreement of Relevance assessments, Irrelevance assessments, and Unclear assessments.

Nine cells have agreement percentages in the 70's (4 fell slightly above the authors' original agreement rate of 73.3%; all others fell below to varying degrees). Eleven cells show agreement percentages in the 60's, 6 cells in the 50's, and 2 cells in the 40's. Again there is noticeable variation in agreement, ranging from a low of 48.3% between Instructors E and H, to a high of 75.8% between Instructors D and G, and G and H, a distance of 27.5 percentage points. The mean agreement of all pairs is 63.9%, which seems a fair or moderate degree of agreement at best.

To summarize, instructors' assessments of response relevance were classified over all students and items, by students' means, by item means, and by the instructors' Relevance assessments themselves and

Table 5. Percentages of Agreement for Each Instructor Pair

Assessor	A	B	C	D	E	F	G	H
A	—	73.3	66.7	75.0	61.6	49.2	72.5	63.3
B		—	72.5	75.0	56.7	57.5	70.8	64.2
C			—	72.5	65.0	52.5	69.2	61.7
D				—	56.7	60.0	75.8	67.5
E					—	50.0	52.5	48.3
F						—	61.7	60.8
G							—	75.8
H								—

Table 6. Summary Table of Results

Item	Relevance	Irrelevance	Unclear
Overall assessments	72.6	18.4	9.0
Range of mean assessments for students 1-10	56.3-82.3	14.6-32.2	3.1-13.5
Range of mean assessment for items 1-12	42.5-97.5	1.3-45.0	1.3-16.3
Range of percentages of Instructors A-H	54.2-85.8	1.0-39.2	0.0-37.7

Note. All assessments measured as percentages.

the degree of their agreement with one another. Three sets of findings emerged, which are summarized in Table 6.

First, instructors found 72.6% of all responses Relevant, and 27.4% either Irrelevant or Unclear. Second, when students' mean relevance assessments were taken, there was a considerable range of variation, 56.3% to 82.3%. This suggests that some students' reasoning for rating selection may be more credible than others. When item mean relevance

Percentages	Frequency
40-50%	2
50-60%	6
60-70%	11
70-80%	9

Figure 1. Comparison of instructor agreement including frequency of agreement in each percentage range. Mean agreement=63.9%

assessments were taken, there was an even greater range of variation, 42.5% to 97.5%. This suggests that some items lend themselves to clearer and more straightforward interpretation by students, and thus more relevant responses. Further, instructors also differ markedly from one another in the percentage of relevance assessments given, a range of 54.2% to 85.8%. This suggests that instructors may define relevance with greater or lesser strictness, or by varying standards. Finally, the percentage of agreement between each possible pair of instructors shows wide variation, clustering between 60-80%, with a mean agreement of 63.9%, reinforcing the perception of variation in instructors' views of relevance (Figure 1).

Discussion and Conclusion

These results show that despite best intentions of an institution, an evaluation instrument with seemingly straightforward and useful items has yielded results reflecting unintended subjectivity and error mixed into what may be useful teacher feedback. It appears that some items ask for student responses with varying degrees of clarity, and students appear variably capable of making meaningful responses. The variability in instructor assessments seemed to further confirm the uncertainty of what constitutes relevant or meaningful student responses. Given the small size of this study, we cannot know how representative these

findings are. Nor is not the purpose of these studies to critique this particular instrument since similar items are so widely used, and in any case the instrument has been updated since the original data collection.

Nonetheless, the combined results of these investigations must surely help to disconfirm the appearance of objectivity attributed to or implied by the many studies correlating SRI data and other measures. Further, the findings of Part 1 and Part 2 show explicitly the possible ways in which SRIs are, as Fich (2003) noted, "low-precision instruments." At least it must be argued, despite nuanced discussions of SRI validity (e. g., Theall, Abrami, & Mets, 2001), that if conventional SRIs and their use are to credibly reflect and inform on teaching effectiveness, either to students or to instructors, they must be amenable to direct demonstration using verbal protocols. For student evaluation instruments to be useful, we must have confidence that these instruments accurately and authentically characterize classroom events and objectives as intended by the instructor, as well as the students' classroom experience, in both item construction and ratings options. Thus, a high proportion of students must find the instrument designers' intended presuppositions, implications, and meaning recognizable, and must be able to respond in kind.

This two-part study illustrates how far afield a conventional SRI may be from these aims. It is conceivable that different items may yield a different result, or that an instrument that has undergone rigorous validation procedures may yield a different result (see Abrami, d'Appolonia, & Cohen, 1990, for a discussion). However, this remains an empirical unknown. Given the fact that many institutions develop their own evaluation instruments, it may be that the unease with conventional SRIs articulated by many critics, which has edged their use away from a central role in faculty evaluation, is well-founded.

Findings of verbal protocol studies such as this one and others mentioned here, and interview studies on instructor perspectives and student characteristics on this topic in the Japanese setting are

significant for tertiary education in Japan and the issue of course evaluation. The acceptance of western-style student ratings instruments for evaluation of college and university classes has spread rapidly. Indeed, it may be that all the features that initially appealed to western administrators are attractive for the same reasons to institutional administrators in Japan. At the same time, alternative approaches to evaluation of faculty effectiveness finding favor in western institutions (e.g. classroom observations, portfolio development) may encounter societal and cultural constraints in Japan, where direct observation and critiquing colleagues' work may seem too intrusive. It remains an empirical question whether SRIs, developed in the context of traditional western academic presuppositions and implications, can be meaningfully employed with valid results in the Japanese tertiary classroom. Abrami, d'Apollonia and Rosenfeld (2007) point out that conventional ratings instruments may not have content validity when used in classrooms employing newer, student-centered teaching paradigms such as cooperative learning (p. 403). Indeed, Burden (2008a, 2008b) reports that instructors in Japan using communicative methods in their language teaching find conventional SRIs feedback lacking relevance and credibility, and unrepresentative of what they do in the classroom. Additionally, Ryan (1998) reviews studies showing that Japanese students judge their instructors differently from students of other countries, for example valuing whether a teacher is creative, entertaining, and approachable over the teacher's content expertise (as in some other countries). Further, studies noted by Ryan (1998) assert that Japanese students evaluate native-speaking instructors and non-native-speaking instructors differently. Whether university instructors in Japan use traditional lectures or not, it remains to be demonstrated how the western model of SRIs constructed in Japan can be made optimally responsive to the Japanese context.

At the same time, given increasing acceptance of SRIs in Japanese institutions, administrators and instructors are in a position to bypass

missteps and errors made elsewhere in countries with longer histories of SRI use. Users in Japan can benefit from expert advice generated over decades, including the recognition of SRIs' imperfections and biases and the necessity of complementary alternative methods of evaluation, however they may be developed in Japan. Further, well-known misuses of results can be avoided, such as comparing individual instructors' means with one another where classroom objectives, content, and student characteristics differ (McKeachie, 1997), or using instrument means rather than global item means (d'Apollonia & Abrami, 1997) or attainment of instructional objectives (McKeachie, 1997; Winskowski & Duggan, 2008) for summative evaluation.

Instructors can additionally take an activist role in their own evaluation of classroom effectiveness, which may include the following:

1. Instructors can become conversant with instructor characteristics that are widely agreed-upon as effective. While we may have reservations about the numerical rating of an instructor characteristic, awareness of such characteristics as organization and clarity can be usefully applied by thoughtful instructors. Long-term work on instructor characteristics by Murray (2007) and Feldman (2007), expertise on college teaching (e.g., Bain, 2004), and discipline-specific experts help point the way. McKeachie (2007) notes that instructors can be trained to employ effective behaviors. Essentially, however, instructors can simply train themselves and each other with methodical observation, feedback, and purposeful innovation, with careful adaptation to the Japanese tertiary context.
2. Instructors can design their own ratings instruments based on the objectives of their course and the activities and events conducted that are intended to accomplish them (Winskowski, 2006; Winskowski & Duggan, 2008). Instructors can incorporate notions of effective teaching into their ratings instruments. For greater usefulness, instructors can implement evaluation before

the end of the semester, when there is still time respond to student feedback. Also helpful is an emphasis on low-inference items which can "be recorded objectively, with little or no judgment or inference, on the part of the observer" (Murray, 2007, p. 184), and which is more informative than high-inference items (Winskowski, 2006).

3. As others have noted, instructors can train students to be critical observers of their own learning and how course events and their own engagement in the course affect that learning. This moves the role of evaluation from a potentially adversarial one and into a cooperative and mutually beneficial one.

Acknowledgements

We are indebted to Bern Mulvey and an anonymous reviewer for helpful comments and advice.

Notes

1. All figures in this report are rounded.
2. Cohen's kappa is a widely accepted indicator of agreement, reflecting how far observed agreement on categorical data departs from expected agreement. While there are no universally agreed-upon standards, a kappa of at least .60 is frequently accepted as "good" agreement.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219-231.

Abrami, P. C., d'Appolonia, S., and Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart

(Eds.), *The scholarship of teaching and learning in higher education: An evidence-based erspective* (pp. 385-456). New York: Springer.

Bain, K. (2004). *What the best college teachers do.* Cambridge: Hard University Press.

Benz, C. R., & Blatt, S. J. (1996). Meanings underlying student ratings of faculty. *Review of Higher Education, 19*(4), 411-433.

Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. (2004). Interpreting course evaluation results: Insight from think aloud interviews with medical students. *Medical Education, 38*(10), 1061 - 1070.

Birnbaum, M. H. (1999). A survey of faculty opinions concerning student evaluations of teaching. *California State University, Fullerton Senate Newsletter, 14*(1). Retrieved March 15, 2005, from http://faculty/fillerton.edu/senatenews/page2.html

Burden, P. (2008a). Does the use of end of semester evaluation forms represent teachers' views of teaching in a tertiary education context in Japan? *Teaching and Teacher Education, 24*, 1463-1475.

Burden, P. (2008b). ELT teacher views on the appropriateness for teacher development of end of semester student evaluation of teaching in a Japanese context. *System, 36*, 478-491.

Cashin, W. E. (September, 1995). IDEA Paper No. 32, Student ratings of teaching: The research revisited. Retrieved Feb. 11, 2009 from http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28*(2), 149-160.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student

ratings of instruction. *American Psychologist, 52,* 1198-1208.

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

Fich, F. E. (2003). Are student evaluations of teaching fair? *Computing Research News, 15*(3). Retrieved March 15, 2005, from http://www.cra.org/CRN/articles/may03/fich.html

Johnson, V. E. (2003). *Grade inflation: A crisis in college education.* New York: Springer-Verlag.

Kolitch, E., & Dean, A. V. (1998). Item 22, "Overall, [the Instructor] was an effective teacher": Multiple meanings and confounding influences. *Journal on Excellence in College Teaching, 9*(2), 119-140.

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

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

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

Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R.P Perry & J.C. Smart (Eds.), *The*

scholarship of teaching and learning in higher education: An evidence-based perspective (pp. 145-183). New York: Springer.

Orsini, J. L. (1986). Halo effects in student evaluations of faculty: A case application. *Journal of Management Education, 101*, 38-45.

Ryan, S. M. (1998). Student evaluation of teachers. *The Language Teacher, 22*(9). Retrieved January 4, 2010, from http://jalt-publications.org/tlt/files/96/sept/learning.html

Selden, P.. (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.) (2001). The student ratings debate: Are they valid? How can we best use them? *New Directions for Teaching and Learning, No. 109*. San Francisco: Jossey-Bass.

Winskowski, C. (2006). Documenting instructor-effectiveness, Part 2: Toward low-inference student ratings instruments. *OnCUE, 14*(1), 9-21.

Winskowski, C. (2010). How students rate, Part 1: A pilot think-aloud study of students' course evaluation responses. *OnCUE Journal, 4*(1), 3-42.

Winskowski, C., & Duggan, S. (October, 2008). Student opinions challenge course evaluations. Japan Association for Language Teaching 34th Annual International Conference, Tokyo. Japan.

Christine Winskowski *holds a doctorate in psychology and a second master's degree in English as a second language. She presently teaches English and area studies at Morioka Junior College, Iwate Prefectural University in Takizawa, Iwate and has taught in the United States and China. Her experience with students' course evaluations spans about 25 years.*

Susan Duggan *holds a master's degree in English education. She has taught French, English, and German, the latter two languages for many years in Iwate. She presently teaches English at Iwate Prefectural University.*

Appendix

Pilot Ratings Study – 2007–08

Assessment: Were the students responsive to the items?

Assessor's Name: _____

Instructions – Please show if each student's response seemed relevant or responsive to the item.

Please the following symbols:

Thank you very much for your help!

O = Completely relevant or mostly relevant

X = Totally or mostly irrelevant

? = Not clearly relevant, but not clearly irrelevant (I am not sure)

	Students									
	1	2	3	4	5	6	7	8	9	10
Item 1– I have had a strong interest in this course from the beginning.										
Item 2 – The content of the lessons corresponded to the syllabus.										
Item 3 – The way of teaching focused on the important points.										
Item 4 – Use of teaching materials was appropriate.										
Item 5 – The quantity of the lesson content was appropriate.										
Item 6 – The teacher's explanations were easy to understand.										
Item 7 – The lesson content was easy to understand.										
Item 8 – The teacher provided opportunities for comments and questions and responded appropriately.										
Item 9 – I was able to sense the teacher's enthusiasm.										
Item 10 – I felt intellectually stimulated by this course.										
Item 11 – I acquired a lot from this course.										
Item 12 – Considering the course as a whole, I feel satisfied with it.										