Feature Article

How Students Rate, Part 1: A Pilot Think-aloud Study of Students' Course Evaluation Responses

Christine Winskowski

Morioka Junior College, Iwate Prefectural University

Abstract

Recently, the use of standardized student course-evaluation forms has become common in Japanese tertiary education. Despite a long history of use in other countries, an extensive number of studies, and many controversies, students' thinking and reasoning processes as they complete course evaluations have received scant attention. In this study, 10 students were asked, while completing a required English language course evaluation form, to respond to two interview questions: 1) What does this item mean? and 2) Why did you give it that rating? Thematic content in the student responses for each of 12 evaluation items are presented in this study, along with the student ratings accompanying each theme. The majority of responses comprise conventional explanations and accounts reporting classroom experience that might be expected and even predicted for a language classroom. At the same time, a striking minority of responses reveal difficulties inherent in many items as well as in student reasoning. Item difficulties include the incorporation of unwarranted or questionable assumptions and premises, and questionable or infelicitous phrasing, ambiguity, etc. Student response difficulties include unexpected and idiosyncratic reasoning motivating rating, and confusion about or confounding of item meaning.

第二言語を学習する場合いくつかの条件がその成功の鍵となる。そのうち多くは感情的なものが占めており、言語学習者の姿勢や信念がもっと注目されるべきである。この論文ではイレーン・ホーウィッツにより発明されたBALLIを使ってタイと日本の大学のEFL学習者にとって有害又は役に立つ信念や姿勢は何かを比較し、その概要を述べている。また、補足として１０４人の大学院生の調査結果もより経験を積み重ね成熟した学生との比較として紹介する。この調査の信頼性を確かめる為に再調査（test-retest）を行い、要素分析（factor analysis）を使って調査結果の編成を行った。

Introduction

The topic of students' end-of-term course evaluations has generated a very large literature worldwide, estimated to number well over 2000 publications (see Feldman, 2007). Recently, the use of end-of-term standardized student course-evaluation forms has rapidly become the norm in a growing number of college and university classrooms in Japan. Following a recommendation by the Ministry of Education (MEXT, 1998) to implement evaluation practices, institutions began adopting the approach widely used in Western and Western-adapted universities. This approach uses a set of 10-20 statements that might describe a university course and its instructor. Students are asked to read and choose from among alternatives, often numbered, which reflect their response to statements describing their experience in the course. So, for example, for "You found the class intellectually challenging and stimulating" a student could respond on a 9-point scale, with alternate numbers labeled "Not applicable –Strongly Disagree – Disagree – Neutral – Agree – Strongly Agree" (Marsh, 1982; see also Winskowski, 2005.)

This is a deceptively simple and attractive design. It is simple because the average class rating for each item can easily be calculated, providing a ready summary of students' ratings. Any instructor can administer this kind of evaluation instrument at the end of the course.

Instructors may find it attractive because it can reveal problems that may not have been noticed (e.g., that the level of lecture or textbook is too difficult). Administrators may find it attractive because it is inexpensive and largely trouble-free to administer and provides information with which to engage the instructors on the effectiveness of their teaching. The design, however, is deceptively problematic, as an enormous volume of controversial literature attests.

The degree of controversy may stem from two elements: On one hand, there is widespread acceptance of conventional high-inference survey-style item construction (Winskowski, 2005) and ratings methodology. On the other, many instructors argue that standardized ratings forms do not represent key characteristics of their classrooms or students' experience of classroom events with reasonable faithfulness. Compounding this is a dearth of direct observation of users' reasoning in ratings selection for item validation. Revealing how students select ratings could allow us to see the ways in which ratings forms do and do not fulfill the intentions of their writers, do and do not address significant elements of teaching, do and do not reflect learning, and do and do not reflect the realities of the classroom. Thus, this study explores "think-aloud" protocols in the process of students' course rating to add to our understanding of these instruments and to point the way to improvements in course evaluation by students.

Literature Review

The preponderance of research literature on student ratings instruments (SRIs) or student evaluations of teaching (SETs) argues that they are reliable and consistently show moderate validity. The best-known contributors use large-scale, quantitative research programs carried out over decades, linking SRIs with student achievement, such as grades and exam results. Well-known examples might include work on multisection validity studies (Abrami, d'Apollonia, & Rosenfield,

2007), and multitrait-multimethod validity studies with the SEEQ ratings instrument (Marsh, 2007), as well as hundreds of correlational and experimental studies supporting this argument. Another key research thread is the identification of teaching dimensions from student ratings instruments (partially summarized in Feldman, 2007), which is especially instrumental in addressing whether teaching effectiveness is a psychological construct comprised of multiple dimensions (Abrami, d'Apollonia, & Rosenfield, 2007), or a single, underlying dimension (Marsh, 2007). The literature additionally identifies predictable biases found in SRI results (e.g., Cashin, 1995; Centra, 1993; Kulik, 2001; Ory, 2001). It is known, for example, that elective courses attain higher ratings than required courses and advanced courses attain higher ratings than introductory courses (Cashin, 1995). Centra (1993) summarizes the findings of two studies on the relationship between student ratings of instruction and course disciplines, one by Feldman (1978) and another by Cashin (1990). Both show that humanities courses generally get the highest ratings; social sciences, education, health professions, English language and a few others get middle-level ratings; natural sciences, engineering, computer science, and other more technical fields tend to get the lowest ratings (Centra, 1993, pp. 67-71). These findings were confirmed in subsequent studies, according to Cashin (1995), and by an extensive analysis by Hoyt and Lee (2002). There is a "halo effect" (Orsini, 1986), where a student's general judgment of the instructor (positive or negative) tends to influence specific assessment of that instructor's course, and a "recency effect" (Dickey & Pearson, 2005), where students give more weight to recent course events than prior events in their ratings, as well as influences from instructor personality characteristics. Finally, the University of Washington, credited with originating SRIs, has begun adjusting SRI medians to "statistically equate classes on students' reason for taking the course, class size, and grading leniency" (Office of Educational Assessment, UW Seattle, 2005), since these are also known and documented biases.

A smaller but strong thread of the SRI discussion details a number of flaws with these instruments and their administration. Scriven (1995), who argues that student ratings can be useful, identifies several problems, including invalid content of the items for use in personnel decisions; errors in administration, report design, and interpretation, among others (p.1-2). A number of writers are more starkly negative about the value of student ratings, including Armstrong, 1998; Birnbaum, 1999; Cruse, 1987; Lewis, 1998; and Trout, 2000. Additionally, this literature reports by instructors that they have manipulated (i.e., "dumbed down") their courses for favorable grades (Stake, 1997); the suspicion that most institutions use home-made rather than validated instruments, since validation is expensive and time-consuming (Abrami, d'Appolonia, & Cohen, 1990); legal challenges to the use of SRIs for administrative assessment of faculty performance rather than improvement of teaching only (Haskell, 1997); evidence that expected grades rather than teaching effectiveness influence student ratings (Johnson, 2002, 2003); evidence that students rate personality rather than teacher effectiveness (Clayson & Sheffet, 2006); and evidence that students confuse an effective course with the professor's expressiveness (Abrami, Leventhal, & Perry, 1982; Williams & Ceci, 1997). Other concerns include the subjective nature of many instrument items and the lack of interpretability of student responses (e.g. what is the difference between ratings of 4 and ratings of 5?); the anonymity of ratings with its consequent lack of evidence, examples, or responsibility for assertions made by raters, among other issues. In recent years, many institutions have added alternative methods of evaluation, including faculty self-evaluation, committee evaluation, use of colleagues' opinions, classroom visits, teaching portfolios, peer review of teaching, and instructor-designed ratings instruments (Seldin, 1999; Winskowski & Duggan, 2007).

The movement away from sole reliance on SRIs reflects some hard-won wisdom gleaned from decades of development and use of

these instruments. Despite this experience, and despite evidence that instructors in Japanese settings have reservations about the value and validity of SRIs (Burden, 2008a, 2008b), many Japanese institutions may be using SRIs alone for evaluation of teaching effectiveness.

Background for this study

A small group of recent studies have used various verbal protocol analyses (Ericsson & Simon, 1993) to investigate students' course evaluation. Verbal protocol analysis uses reports of reasoning, memory searches, and other mental processes to understand reading, mathematical calculations, and similar cognitive operations. Billings-Gagliardi, Barrett, & Mazor (2004) conducted "think-aloud interviews" while medical students completed a science course evaluation form. They found students based their judgments on "unexpected" (i.e., unique or idiosyncratic) criteria (e.g., their own ability or caring or effort of the faculty member). Kolitch and Dean (1998), in a study of student ratings on the global item "the instructor was an effective teacher," found students used a variety of ways to rate, including finding an average of prior ratings to rate overall satisfaction, responding to feelings and emotions, and identifying a critical dimension by which to evaluate. Benz and Blatt (1996) asked students to respond in writing to the question, "Why did you rate this item as you did?" Identifying themes in the content of student responses, they found that students had various interpretations of items, offered various kinds of evidence for their choices, and had various assumptions about teaching.

These studies offer a significant challenge to the claim of SRI validity, which has in large part been demonstrated by the correlations between SRIs and measures of student achievement (exam results and grades), as well as peer ratings and instructor self-ratings (see Aleamoni, 1999; Cashin, 1995; Johnson, 2002). While these correlations are stable, they do not reveal from direct observation that students' reasoning

in the process of rating addresses recognizable elements of teaching effectiveness. It must be demonstrated that students are expressing their satisfaction with a recognizable, consensually held understanding of the construct "teaching effectiveness." Otherwise, the claim of construct validity of SRIs, and other forms of validity in the ratings process, fails.

The purpose of this investigation is to contribute to a fundamental understanding of the nature of students' evaluation ratings data, particularly in view of the recent use of student ratings forms in Japanese universities (Ruthven-Stuart, 2004). This two-part study is an initial attempt to understand what students are thinking and reasoning when they fill out course ratings forms, and how relevant to the item this thinking/reasoning is judged by instructors. Part 1 reports on the results of interviews conducted with students, asking them to explain their reasoning aloud, as they filled out a course rating form used for evaluation. In Part 2 of the study (forthcoming), a small number of teachers were asked to judge whether these students' interview responses appeared relevant to the item, irrelevant, or indeterminate.

Method

Students were interviewed as they completed a 12-item SRI for a foreign language course to determine their understanding of the evaluation items and the reasons for their rating selections. The numerical characteristics of the students' ratings were first examined. Then from the interview transcriptions, patterns of content in the students' interview responses to each item were analyzed for thematically similar or characteristic content.

Participants

Ten students in a second-year EFL course volunteered to participate in this study. The EFL course was part of a four-semester English

requirement in a college international studies program, located in northern Japan. The students volunteered in their third semester. The course, taught by an instructor with qualifications in TESL/TESFL, focused on listening and speaking skills.

Materials

A conventional, mandatory 12-item student ratings instrument, designed by the institution and administered in Japanese, was used (see Appendix). Each item was a statement with one of the following six Likert-style choices: This statement is: 1 = … not appropriate (あてはまらない); 2 = …is mainly not appropriate (ややあてはまらない); 3 = …is more inappropriate than appropriate (どちらかと言えばあてはまらない); 4 = …is more appropriate than inappropriate (どちらと言えばあてはまる); 5 = …is mainly appropriate (ややあてはまる); 6 = …is appropriate (あてはまる).

Procedure

The students received a group orientation where the purpose and procedure of the study were explained in Japanese by the interviewer, and interview appointments were made. Students then met individually with the interviewer, and filled out the evaluation form in Japanese for their English class while being interviewed in Japanese. For each item, the interviewer asked the student two questions: 1) What does this item mean? 2) Why did you give it the rating that you did? The interviewer sometimes followed up with prompts for clarification. The interviews were recorded, transcribed, and translated into English for analysis by the interviewer.

The ratings data that students provided as part of their interviews were first examined to see the distribution of ratings and their characteristics. These were then compared with the distribution of ratings from the official, end-of-semester student course evaluations to determine how representative the pilot interview sample was.

The students' transcribed responses to the interview questions for each item were then examined. Many times, students' responses to a given item expressed thematically the same or similar ideas, comprising descriptions of events, observations, comments, and reasoning. These themes in the response content were especially noted when referred to by two or more students. Representative phrases and statements reflecting each theme, together with the student's ratings, were classified together and titled. Observations were then drawn from these thematically grouped responses.

Results

Distribution of students' ratings

The evaluations given by students in this pilot study were generally quite high. Table 1 shows how frequently the 10 students chose a given rating for Items 1-12, which are shown in abbreviated form.

Table 1

Distribution of Student Ratings of 1–6 for Items 1-12, and Total Percentage of Responses for Each Rating Choice

Rating choices: This statement is...	1	2	3	4	5	6
1. I had prior interest in this subject.	0	0	0	4	4	2
2. Content corresponded to syllabus.	0	0	0	2	3	5
3. We focused on important points.	0	0	0	0	1	9
4. Teaching materials were appropriate.	0	0	1	1	2	6
5. Quantity of lesson was appropriate.	0	0	2	0	4	4

6.Explanations were understandable.	0	0	0	0	1	9
7. Content was understandable.	0	0	0	0	3	7
8. I could give comments and questions.	0	0	0	0	1	9
9. Teacher was enthusiastic.	0	0	0	0	0	10
10. I was intellectually stimulated.	0	0	0	1	2	7
11. I acquired a lot.	0	0	0	1	0	9
12. I was stimulated.	0	0	0	1	3	6
Totals	0	0	3	10	24	83
(% of response)	(0)	(0)	(2.5)	(8.3)	(20)	(69.2)

Note. Student distribution numbers are raw totals.

It can be seen in Table 1 that ratings of 6 comprise over 69% of the total. Another 20% of ratings consisted of 5s and a total of 97% fell into the upper half of the ratings range, 4-6. Certainly, students' ratings in this pilot study were strongly positive.

A comparison of the pilot ratings totals in Table 1 can be made with the official, year-end ratings from the instructor's combined sections (*N* = 28), from which the 10 pilot students were drawn, shown in Table 2. This allows us to see whether, and to what extent, the pilot ratings departed from the year-end ratings.

Table 2

Distribution of official, year-end student course rating totals of 1-6 for Items 1-12

Rating choices: This statement is...	1	2	3	4	5	6
Totals	0	1	4	33	96	201
(% of response)	(0)	(3)	(1.2)	(9.8)	(28.6)	(59.8)

In Table 2, it can be seen that there are nearly 10% fewer ratings of a 6 than in Table 1, and about 8.5% more ratings of 5, suggesting a slightly higher profile of ratings for our pilot group. This higher profile may be due to random variation or to a social desirability effect (i.e., that students responded in a way that they imagined would be approved by the interviewer or others). Apart from these differences, the overall distribution of the pilot ratings and official ratings seem similar.

Student interpretations of items and their rationales

For the initial three to four items, students' paraphrases of the item meanings tended to also be the reasons that they gave a particular rating (i.e., that the instructor conformed to a description of the item). As the students completed Items 4-12, they tended to respond more distinctly to each question.

Results of students' interview responses are provided for each of Items 1-12 separately. The students' spoken responses for each item are classified by themes, most reflecting consistencies or similarities in the content of two or more responses, including reference to similar or related events, items used in class, claims, reasoning and arguments. For example, for Item 1. I had prior interest in this subject, three themes in the students responses emerge: Theme 1 - What happened in class, for example, expressed as "my level of English improved" and "it was enjoyable"; Theme 2 - Interest before class started, for example, "I've always been interested"; and Theme 3 - Compulsory status of the class, for example, "the course was compulsory."

The responses classified under each theme are shown with representative phrases designed to illustrate their range and substance. Students' ratings of 1 to 6 follow each phrase; a phrase with two or more ratings (e.g., Rating 4, 4, 5) indicates that two or more students mentioned essentially the same content. Additionally, of course, a single student response could include comments classified under more than one theme. Thus, elements of a student's response might be

distributed over themes. For example, Item 1 shows the 10 students' responses yielded 13 representative phrases classified into 3 themes. Some students said virtually the same thing in two cases; thus, there are two mentions of "liking English a bit" (ratings of 6 and 4), and six mentions of the course being "compulsory" (ratings of 4, 4, 4, 4, 5, and 5). Occasionally, a single response fits into no identifiable theme, and so is considered a separate theme. Finally, for each item and set of response results, observations are made on how the students responded and what their responses reveal.

Item 1. I had prior interest in this subject

(この授業にはもともと強い関心があった.)

Students' explanations of the item meaning were straightforward (e.g., "I took it as, 'Were you interested in this course?'") and are not reported separately from reason for rating. Students' reasons for rating fell into three themes. Theme 1 included 3 references to what happened to students in the class, presumably reflecting interest in the course. Theme 2 shows 9 mentions of interest, partial interest, or lack of interest in the course. In Theme 3 there were 7 spontaneous mentions of the fact that the course was compulsory, one student saying s/he disliked the compulsory aspect. Also, strength of ratings seemed to correspond with two themes, What Happened in Class (ratings of 5-6), and Compulsory Status (ratings of 4-5), but not entirely with Interest (ratings ranging from 4-6).

Theme 1 – What happened in class
My level of English improved – Rating 5
It was enjoyable - 6
We did practical English, could have conversation with friends - 6

Theme 2 – Interest before class started
I've always been interested – Rating 5
I'm interested - 6

I'm not aware of interest – 4
I'm interested, but not really - 4
I'm not very interested – 4
I like English a bit – 6, 4
I wanted to do English – 5
I'm interested, but more so in other subjects - 5

Theme 3 – Compulsory status of the course
The course was compulsory – Rating 4, 4, 4, 4, 5, 5
I disliked the compulsory aspect - 4

Results of Item 1

The responses to this item offer evidence of a bias noted in the literature, namely that required courses get lower ratings (Aleamoni, 1999; Cashin, 1995), although this finding is not known to be confirmed with other Japanese student data at this writing. This is the only item whose response themes are so distinctly ranked by rating, with a preponderance of lower rankings associated with mention of compulsory status. However, the ratings form allows no indication of the course's required/elective status, and if administrators or faculty members are not aware of this potential bias, they may misconstrue ratings data. A student response to this item may in no way reflect teacher effectiveness, but rather student characteristics. In fact, other student ratings should probably be considered in the context of students' responses to this item, since other ratings may be influenced by students' interest in the course.

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

(授業内容がシラバスに沿っていた.)

Students' explanations of item meaning were again straightforward and are not reported separately. Reasons for their ratings clustered in four themes. In Theme 1, 6 students named course elements which presumably were in the syllabus. Theme 2 has two references to the

compulsory status of the course, which do not seem relevant to the item but were perhaps still on the minds of the students after discussing Item 1.

Students indicated if they knew the syllabus or not, resulting in responses in Themes 3 and 4. Five students said they had "not really" or had "more or less" looked at the syllabus, or that the course seemed "the same as" prior courses in the series. One student asked the interviewer for a copy of the syllabus.

Theme 1 - Lesson elements corresponding to the syllabus
We used a textbook, did communication, watched movies – Rating 6
The course is based on textbook, we did games, watched videos - 5
We focus on speaking, listening; we watched videos, did homework - 6
We used a textbook - 4
We did not only writing, but also listening, speaking, related activities - 6
We did not only grammar, but also conversation - 6
Theme 2 – Compulsory status of the course
I disliked compulsory aspect – Rating 5
The course was compulsory - 5

Theme 3– Students' reading of the syllabus
I haven't really looked; I don't know, maybe 4 or 5 – Rating 5
I haven't looked at the syllabus but I'm satisfied, so it's "5" - 5
I haven't looked much; didn't look – 4, 5, 6
I did look at the syllabus – 6, 6,
I more or less know what's in the syllabus [i.e. it is continued from first year class] – 4, 6
Do you have a copy of the syllabus? - 6

Theme 4 – Comparisons of 1st-year and 2nd-year
This course was the same as first year – Rating 4, 5, 6

Results of Item 2

Since students were familiar with the course from previous semesters, it seems unsurprising that 5 students did not look at the

syllabus, including the student who asked the interviewer for a copy of the syllabus. How this item reflects effective teaching is unclear. Certainly, delivering a course related to the syllabus is usually a desirable practice, yet there are many situational variables that must be clarified before we can know if the correspondence of syllabus and course content reflects teaching quality. If syllabi are published once a year, as is practiced in some institutions, they may deliberately be left vague to allow the instructor the option of changing/refining plans. Instructors may or may not provide a more detailed course syllabus at the beginning of class. This instructor did so. On the SRI in question, there is no place to indicate whether a student has seen the syllabus, or whether there is more than one. As we see above, students may feel justified in not reading the syllabus, but they will answer the item anyway. Finally, instructors depart from a syllabus for many legitimate reasons. If students respond literally to the item, it could penalize the instructor.

Item 3. The teaching methodology focused on the important points.

(教え方は要所をおさえていた)

Two themes emerged from the responses. Theme 1 showed how students characterize important points. Four students invoked the textbook, syllabus, or "necessary things," suggesting a "learning from authority" model of pedagogy (vs. a functional, communicative, or pragmatic approach). Theme 2 identified teacher or student actions in the classroom, which seemed to indicate important points, although the claim that a teacher explained in "an easy way" may or may not be an indication of "a focus on the important points."

Theme 1 – How "important points" are characterized
Did the T[eacher] focus on [important points]? – Rating 6
Did the course follow the textbook - 6, 6 Important points in the syllabus – 6
Did she teach the necessary parts/things? – 6, 6

Did T do her best? - 6
The things we did every time - 6
Lessons were designed to emphasize conversation - 6
Sometimes I wondered what was the main point – grammar or practical English? – 5
How we progressed, smoothness, whether pointless things done, how we were taught, how we felt, when we mastered how to do something – 6
Important points has to do with the syllabus…oh, but that's Item 2, isn't it? – 6

Theme 2 – Teacher's focus on "important points"
When students didn't understand, T explained in an easy way, with details – Rating 6, 6
T repeated activities, followed textbook – 6, 6
We did conversation, it was practical, did other things – 6
We really followed the textbook - 6
She wanted us to remember – 6

Results of Item 3

It should be noted that two students could not answer this item immediately, and went on to other items before coming back to it. One explanation may be that this item seems more suitable for a content course, rather than a skills course. We may reasonably ask of a history, biology, or sociology course whether important conceptual points were distinguished from the many details that make up the course. While key concepts may be raised in a language class, such a class is ordinarily focused on the process of acquiring, practicing, and producing a large number of discrete vocabulary, grammar, and sociocultural elements with the ultimate intention of integrating them into complex language performance. Thus, if students are attempting to respond to an item that poorly characterizes the nature of their classwork, the validity of their ratings may certainly be in doubt.

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

(教材（黒板、視聴覚教材、テキスト、配布資料など）の使い方は適切であった.)

Responses in Theme 1 describe various effects of materials. These do not directly respond to the interview questions, but seem to express approval of the materials. Theme 2 identifies various meanings that the students saw in the item, such as citing appropriate use of the textbook, the board, and explanations of materials. In Theme 3, students describe various ways materials were used as their reasons for rating.

Theme 1 - Effect of teaching materials used

Photos, colored copies were used, it was fun and made studying easy – Rating 4

Today T made her own handout, it was easy to understand, new and interesting - 6

Theme 2 – Meaning of the item

Textbook, visual things – Rating 6

"The use" was like "the way of teaching" - 6

Did you use the textbook properly? - 5

Were things used in class explained properly so we could incorporate them into lessons? – 6

Was the blackboard used not just for notes? Could you understand the board? – 6

Were the things used well-matched according to situation; did T write on the board if necessary? – 3

Theme 3 - Reason for rating

[We used] not just the textbook; T prepared topics, handouts – Rating 6

Whiteboard use was good but video use, listening on computer, listening on CDs at home would be better – 4

T sometimes erased board with hand, so it's a "5" – 5

Things were properly explained – 6

Textbook was too easy – 3

I'm satisfied with teaching materials - 6

The textbook was not used every lesson (i.e., other materials also used), so [I partly agree] – 5

There were quizzes about videos, we used handouts while doing fun things - 6
T used the board and textbook properly - 6
Apart from the textbook, teaching materials were handouts & things made, "props," and they were fun - 6

Results of Item 4

Students approved of most ways that materials were used, and had reservations about some. These are informative comments for an instructor. Curiously, one student expressed disapproval of the teacher erasing the board with her hand, citing this as a reason for giving a rating of 5; yet this seems a matter of social, not pedagogical, appropriateness. Another student gave a 5 because the textbook was not used in every class. These comments illustrate the sort of potentially unexpected perspective on the students' part, which may remain completely hidden by the neutral appearance of a numerical rating. Similar findings were observed in Billings-Gagliardi, Barrett, and Mazor (2004).

It is possible that students can recognize serious misuse of materials, but may not understand pedagogical intent in the way that an instructor does, for example in the use or nonuse of the textbook. Thus, this item may have some validity, but low precision. The item might generate more helpful responses if it were accompanied by examples and illustrations, or even some training of the students to recognize what constitutes "appropriate use" of materials as determined by the discipline and course type (lecture, practicum, seminar, etc.). However, students' ability to respond may also depend on their age and experience.

Item 5. The quantity of the lesson content was appropriate

(授業内容は量的に適切であった.)

Theme 1 describes the item meaning, linking pace, content, and the amount of class time, sometimes expressed as a question. Theme 2 shows the reasons students chose their ratings. Several felt the pace

or amount of work was slow, some felt it was fine, and some felt the time was short.

Theme 1 – Meaning of the item

If the pace of the lessons is too fast, if you do a lot of things in a short time… - Rating 5

During class time was there anything pointless? Was the course directed at the students as it progressed? – 6

Given the time, we didn't just rush; did you go at an appropriate pace or not? – 5

Was the amount easy for you to do? – 6

Was the lesson content exactly right for the time period? - 3

Theme 2 - Reasons for rating

It would have been better to do more than one point per class - Rating 3

We accomplished something every time, but there were times when progress was slow – 5

Occasionally there wasn't enough time to finish, maybe content should have been reduced - 5

Lessons seemed short - 5, 6

I wanted the pace to be faster, so I put "5"; wanted faster progress – 5, 5

T kept to time well and proceeded appropriately, so I put "6" - 6

There wasn't enough time for everything, we ran over [the class time] 2-3 minutes - 5

The homework was easy to do - 6

It [the item] is similar to Item 4; it [class] could have been more difficult - 3

The lessons are 45 minutes, short, not boring, good for concentrating and learning - 6

Results of Item 5

All these explanations for ratings' choices seem responsive to the item and informative about the course and the instructor's delivery of the course. However, one student referred to the homework, raising the question, Is the homework to be regarded as part of the lesson?

Item 6. The teacher's way of speaking was easy to understand

(教員の話し方はわかりやすかった.)

In Theme 1, students listed the many ways that the teacher made her explanations easy to understand, incorporating explanations of the item meaning as well as reasons for rating. Additionally, all but one rating for these elements are 6 (This statement is appropriate), showing a strongly unified student perception. Theme 2 comprises paraphrases of the item. Item 6 is ambiguous in Japanese, as well as in English. It may be seen as asking whether the content of the instructor's speech is easy to understand, or about the clarity of pronunciation and volume (i.e., elocution) of the instructor's speech. It is not clear what was intended, and responses reflecting both interpretations can be found.

Theme 1 - How the teacher made it easy to understand/reasons for rating

When we didn't understand, the T spoke slowly, explained in a different way – Rating 6, 6

The T explained in simple English so it was easy to understand – 6

The T explained carefully, occasionally used Japanese, explained in detail – 6, 6

The T explained in a different way – 6,

When we didn't understand, the T used a dictionary, put it a different way, spoke slowly - 6

When we don't understand, T mixes Japanese into her explanations – 6,

The T explained how to do the practice, it was easy to understand – 6

It was interesting – 6

The T used Japanese when we didn't understand and she was good at explaining- 6

The T repeated when we didn't understand – 6

The T's pronunciation was easy to understand; [but] because it's in English, I couldn't understand everything – 5

Theme 2 - What the item means

It's the way of speaking, the way the lesson progresses, not being boring – Rating 6

This means the T's explanation of the lesson content and things, was pronunciation easy to understand? - 5

Was the way the course progressed easy for you to understand or not? - 6

Results of Item 6

The one student gave a rating of 5 (This statement is mainly appropriate) rather than 6 because the instructor's explanations were in English and she/he could not understand everything – apparently taking a literal reading of the item. Thus, it seems this response was a reflection of the student's language ability rather than the teacher's effectiveness at explanation. Would this be an acceptable rendering to the item writers? Similar findings, where students have rated an item based on their own ability rather than teacher or course merit, have been found by others (see Billings-Gagliardi, Barrett & Mazor, 2004; Burden, 2008b).

Despite high ratings awarded to this item and students' responses, there is an item bias in the item wording toward "easy" instructor explanations. Hadley & Yoshioka Hadley (1996) and Ryan (1998) have argued that Japanese university students may have a stronger inclination to consider "understandable" and "easy to understand" qualities of a good teacher than students in other countries. However, might instructors sometimes need to explain legitimately difficult material? A better item might ask, "Are the teacher's explanations clear and understandable?" thus avoiding confounding success of the teacher's explanation with easiness, language used, or pronunciation.

Item 7. The lesson content was easy to understand

(授業内容はわかりやすかった.)

As was noted by some students, this item is phrased in a similar way to Item 6 but is focused on the course content. One student asserted that the instructor's explanations were easy to understand, before being prompted by the interviewer to focus on lesson content.

Theme 1 shows responses describing the way in which lesson content was easy to understand as the reason for rating. Students mention the classwork, the instructor's actions, and what they were able to accomplish, among other things. As with Item 6, these responses

raise the issue of what was meant by content. Theme 2 reflects students' definition of the item. Most descriptions referred to the content of materials or the activities.

Theme 1 - Lesson content's ease of understand/reasons for rating

We did the textbook, grammar – Rating 6

My answer is from things used during class, textbook and handouts - 6

The explanations were careful, easy to understand – 6, 5

The examples were short, topics were familiar, we could speak in English – 6

We were able to do things for ourselves, e.g. watch videos, getting explanation in class; I understood it very well – 6

Grammar and vocabulary was emphasized – easy to understand – 6

It was in English, so easy to understand, but sometimes I couldn't understand everything – 5

Known and unknown materials were mixed, so difficult things were made easy – 5

When we didn't understand, T[eacher] used different explanations – 5

We did the same things as in junior high school and high school – 6

Sometimes I couldn't understand lesson content or the English vocabulary – 5

Some courses are hard to understand, but English is easy to understand, [but] I couldn't understand everything – 5

Theme 2 - Definition of" lesson content"

[It's] textbook, grammar, past tense ; handouts and videos – Rating 6

[It's] content of the textbook – 6

[It's] watching videos, textbook, new materials, practicing English – 5

Lesson content is what you really do, making groups, speaking practice, little tests – 6

Did you understand what you were studying during the lesson or not? – 5

Were the things you did simple? - 6

This is the same as Item 6 – 6, 6, 6

Results of Item 7

As with Item 6, two students gave a "5" (see Theme 1) because they could not understand all of the English that was used, that is, rating according to their own ability, rather than the instructor or course.

Generally, students seemed to understand this item and respond in a way that is similar to the responses in Item 6, reflecting successfully accessible content in this course. Additionally, as with Item 6, the question is raised whether "easy" lesson content is always desirable lesson content. Ryan (1998) concludes from a survey of Japanese and Australian students that

> …Japanese students are much less concerned about the subject-mastery of their teachers than are Australian students. Instead they are eager to have a teacher with a wealth of knowledge about life in general, a fund of jokes and funny stories, and wisdom in the art of teaching. (p. 3)

Nonetheless, course content may be legitimately difficult. If students do not understand this, instructors could conceivably be penalized for delivering challenging material at what they consider to be at a correct level of sophistication for college or university students.

Item 8. The teacher provided opportunities for comments and questions and responded appropriately

(教員は、発言や質問の機会を設け、適切に対応していた.)

Theme 1 includes ways in which the instructor provided opportunities for comments and questions, both explaining the item and providing the reason for rating. The students seemed to emphatically feel they had opportunities to ask questions, with one reservation in Theme 2 expressed due to the 45-minute time limit. Here, as elsewhere in the data, some responses were expressed as questions.

Theme 1 – Definition of item/Teacher's provision of such opportunities

The T provided opportunities for questions; when we didn't understand, T explained/joined in until we did – Rating 6, 6

The T made students talk a lot; I think she made everybody talk - 6

The T always asked us, "Do you have any questions?" – 5

Time was provided for students to give ideas and ask questions – 6

When we didn't understand, the T came and explained to us – 6

The T made sure to ask "Who knows the meaning of this sentence?" - 6

It wasn't only the T talking, she let students ask questions; Did she communicate with us? Did she answer our questions properly? – 6

The T listened and asked if we understood – 6

The T often asked if there is anything we don't understand – 6

Did she make lessons that were easy to understand? - 6

Theme 2 - Time issue

Sometimes, with a half [period]-class the time becomes short; there was not enough time – Rating 5

Results of Item 8

Students' responses to this item seem unproblematic. It can be noted that this item asks for a response that relies on students' direct reporting on their own experience and their observation of the class events. Thus, student answers seem directly relevant to the interview questions.

Item 9. I was able to sense the teacher's enthusiasm

(教員の熱意が感じられた.)

In Theme 1, students described ways in which the instructor indicated enthusiasm, including classroom efforts, preparation, attitude, and even the instructor's activities outside the class. In Theme 2, students explained their understanding of the item. Two students refer to feeling (feeling the style of teaching, feeling the enthusiasm), whereas others referred to the teacher trying her best or trying hard, motivation, careful explanation, and teaching in an interesting way. One student had

difficulty identifying what this item meant, and another responded as if Item 9 were equivalent to Item 6 ("The teacher's explanations were easy to understand").

Results of Item 9

Theme 1 – Teacher's indication of enthusiasm/reason for rating

Lessons were easy to understand; T explained until we understood - Rating 6

The T does TOEIC and volunteers at the International Exchange Center – 6

The T did different things, brought interesting food to try– 6

The T's attitude in different lessons, prepared different things for us, prepared a handout every time, suggestions for ways to study – 6

There is no time when she is not enthusiastic - 6

The T tried hard to do easy-to-understand lessons – 6, 6,

The T learns Japanese enthusiastically – 6

The T always thought about what we were going to do, activities were always different, so the lesson was fun - 6

Theme 2 – Definition of the item

The feeling of the style of teaching – Rating 6

Maybe it means, "Did the T[eacher] try her best?" – 6, 6

Every time, the T explained carefully, taught in an interesting way, really tried hard – 6

Enthusiasm…enthusiasm…it's a bit difficult to understand; in any lesson any T tries hard to teach, I think…I suppose – 6

[It means] Do you feel the enthusiasm of the T? – 6

Whether the T was motivated; maybe it's the same as Item 6 – 6, 6

Theme 3 - Comparison to other teachers

I'm not comparing the T with others – Rating 6

I compare the T with other Ts - 6

For this item, the interviewer spontaneously asked Students 6 and 7 whether they compared this instructor to other instructors, or only considered the instructor's teaching individually. One student did compare the instructor to others, and the other did not. While this question was not asked of all students, it raises the two questions: 1) Did the ratings sheet makers intend students to compare their instructors? and 2) Does comparing or not comparing impact ratings, and if so, how? If it does, the impact on evaluation should be investigated and taken into consideration. Felder (1993) incorporated comparison into his item design so that students rated their instructor as one of the top 3 or 4 instructors, one of the top 25%, one of the middle 40-75%, and so on.

Despite the high praise this instructor garnered, reservations can be raised about this kind of item, which essentially asks students to speculate on an instructor's psychological state. First, it invites an inappropriate degree of subjective imagining by the student. It is preferable to ask students to restrict their reporting to observable events. Second, university instructors do not always have the luxury of teaching only subjects they feel enthusiastic about, and must accept their share of service and other courses. While a number of institutions prefer instructors with credentials in TESL/TEFL, it could be argued that there are many university instructors of foreign languages in Japan whose specialties are in literature, philosophy, history, and various social science fields, some of whom almost certainly have marginal interest in language teaching. Next, the item presupposes that good teaching is enthusiastic teaching. However, might a reserved instructor not also be effective? This item may penalize a reserved instructor who is in fact effective, but not extroverted. Some research does show that enthusiasm, expressiveness, and other forms of "educational seduction" (Naftulin, Ware, & Donnelly, 1973) resulted in higher student ratings regardless of the level (high or low) of information delivered in a class (see also Clayson & Sheffet, 2006; Williams & Ceci, 1997; Williams

& Ware, 1977). However, enthusiasm had only minor impact on student achievement (Abrami, Leventhal, & Perry, 1982). While some of this research has been criticized on methodological grounds, and while personality traits such as enthusiasm and lively performance will certainly be enjoyable for students, it is best not to confound enthusiasm with teaching effectiveness.

Item 10. I felt intellectually stimulated by the course

(この授業によって知的刺激を受けた.)

Students' explanations of class events indicating intellectual stimulation were in fact the reasons for their ratings. These were separated (Theme 1) from statements that seemed to be a direct definition of "intellectually stimulated" (Theme 2):

Results of Item 10

Theme 1 – Indicators of intellectual stimulation, reason for rating

I learned things I didn't know – Rating 6, 6
We learned American customs – 6
I didn't have conversation lessons before – 6
The T's course was fun – 6
It was interesting – 6
I gained a lot of knowledge – 5, 6
I couldn't remember it all – 5
After class, enthusiasm dropped – 5
We learned different ways to say things and old, polite expressions – 6
[There were] occasionally new words, but most content was already known – 4
We did speaking lessons, so I improved ability, was motivated to speak, listen to music, read – 6
[We] learned many new things, different things, practical expressions – 6
We could answer questions, increased our vocabulary, express feelings better – 5
[We] learned information about various countries – 6

Theme 2 - Meaning of intellectual stimulation
Did you gain a lot of knowledge in this course? – Rating 5
It's similar to [Item] 11. Did we learn how to do things? – 6
That's difficult…is what you learned useful? Do you remember what you learned? – 6
Did your vocabulary increase? – 5
Did you learn a lot? – 4
Do you remember what you learned? – 6

In the first theme, students referred to learning new things and to improving ability and knowledge. It seems that most students may have meant, "It changed what I could do; I learned something." Many student responses seem really to be answering Item 11, that is, what they learned from the course, including new information, American customs, different expressions, and vocabulary, as well as mentioning increased ability. Responses of the second theme, with reference to gaining knowledge, remembering what was learned, learning how to do things, and increasing vocabulary, also seem to be the intended focus of Item 11.

"Intellectual stimulation" may be a difficult concept for many to define or explain, and student responses here suggest at least some confusion between what was learned and the stimulating effects of that learning on their thinking and reasoning.

Item 11. I acquired a lot from this course

(この授業によって得たものは多かった.)

The data show that students' item explanations and reasons for rating cluster in the first theme, all describing examples of what was acquired, such as thinking in English; learning about culture, grammar, vocabulary, customs; skill improvements, increased motivation and confidence, and others. The second theme shows that 4 students asserted that Item 11 was similar to, or the same as, Item 10.

Theme 1 – Indicators of acquiring a lot

I thought in English, thought about grammar, learned about American culture – Rating 6

I learned not only from lessons, but also about activities at International Exchange Center – 6

I became more confident, now do volunteer activity– 6

Now I'm motivated to speak to native [English] speakers; listening ability improved – 6

I acquired English ability, customs, food, fashion, overall knowledge – 6

Knowledge, ways of thinking, practice – it was all beneficial – 6

I had listening practice; learned a lot of words and things – 6

I acquired many things from every lesson – 6

The T[eacher] answered questions at normal speed; vocabulary increased; became able to talk and communicate – 6

We were taught a lot of words for daily conversation; they were easy to adopt; the kinds of words suitable for a foreigner – 6

Knowledge of English words and things increased, doing conversation, conversational expressions – 4

Theme 2 – Meaning, mention of Item 10

The same applies as for Item 10 – Rating 6

This is similar to Item 10; do you feel you have learned something? – 6

The meaning is the same as for Item 10 – 6

It's kind of the same as Item 10; reason is the same [I chose 4 because most of content was known] - 4

Results of Item 11

This item appears essentially unproblematic, as students indicated a number of things they acquired and how they acquired them. However, mention of the similarity between Items 10 and 11 suggest that the two items were not distinct in at least some students' minds, and again raises the issue of item validity.

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

(総合的に考えてこの授業に満足できる.)

The majority of student responses to Item 12, shown in Theme 1, comprised reasons for being satisfied or not satisfied, sometimes by way of explaining the meaning. Students frequently named class activities they favored here, such as using a variety of materials, being active, and having interesting lessons. They also express reservations or regrets about some events; they wanted to do more, wanted more chances to speak in pairs, and didn't like the 45-minute time periods. Theme 2 reflects how students defined satisfaction. While one referred to a sense of accomplishment, three students indicated that their measure of satisfaction was based on a rough average of their ratings of the 11 prior items.

Theme 1 – Reasons for satisfaction/lack of satisfaction
We used a variety of materials – Rating 6
It was really fun – 6 , 6
We studied a lot of English – 6
We read stories – 6
We were active – 6
I was motivated – 6
My English ability – 6
I looked forward to class – 6
The T's enthusiasm was great – 5
I wanted more chances to speak in pairs – 5
I enjoyed communicating – 6
I didn't like the 45-minute time period, too fast – 5
I learned a lot, but our level was low ; need a better textbook – 4
Things learned included content knowledge, English use – 6
Lessons were interesting – it was good to talk about ourselves in English – 6
We did unusual things, e.g. watching videos, DVDs, movies – 6
I acquired a lot, but wanted to do more; we could have progressed more – 5
My level now is lower than when in high school - 4

Theme 2 - Definition of "satisfied"
A sense of accomplishment – Rating 6

I always take the average of No. 1-No.11 - 6
Based on all the questions up to now, I was satisfied - 6
I check through the numbers…I take the average – 6
Looking at the first term [and] as a whole, as at this list of questions, were you satisfied? - 4

Results of Item 12

This type of item, commonly called a global item (Kolitch and Dean, 1998), is often recommended for summative evaluation, and as such, the students' responses about satisfaction or dissatisfaction with the course are informative. At the same time, it is of interest that 1 student reports rating holistically, where 3 others determine their rating from the average of prior ratings. This difference in determining the rating is a finding observed by Kolitch and Dean (1998). It is an empirical question whether the two approaches impact ratings differently, adding to the pool of unknowns about how students rate their courses.

The elements of students' interviews amenable to classification into themes appear largely responsive to the evaluation items as they are constructed. The majority of responses comprise conventional explanations and accounts reporting classroom experience that might be expected and even predicted for a language classroom of this type. At the same time, a significant minority of these responses reveal difficulties inherent in many items as well as in student reasoning. Item difficulties include the incorporation of unwarranted or questionable assumptions and premises, questionable or infelicitous phrasing, ambiguity, etc. Student response difficulties include unexpected and idiosyncratic reasoning motivating rating, and confusion about or confounding of item meaning.

The difficulties noted with these items and responses are not atypical, and it should be noted that the college administering this SRI updated and revised its rating instrument shortly after this data was obtained.

Summary and Discussion

In this study, students provided interview responses on their understanding of item meaning and reason for rating as they completed a conventional, 12-item course evaluation form. While the majority of student responses seemed sensible, reasonable, and even predictable, a significant minority of responses reveal problems in item construction and student responses. These specific vulnerabilities indicate the vulnerability of such instruments as a whole. Rating items should characterize classroom events in a common sense and recognizable way to students, and must readily capture their experience of the teaching and how students were engaged with the teaching. If this basic criterion is not fulfilled, an SRI cannot then have construct validity, embodying a sound theoretical understanding of teaching effectiveness, clearly operationalized in its items with measurable indicators.

The data obtained here offer valuable insights and may explain the dichotomy of views on SRIs: The ways in which evaluation items and their responses are problematic may be masked to those who advocate standardized evaluation instruments based on stable correlations to related measures such as grades, peer evaluations, and self-evaluation. Indeed, often the unit of analysis for large-scale validity studies is the mean class rating, which would certainly not reveal the difficulties with item validity illustrated in this study. Further, instructors in the classroom may be aware that students are responding to problematic ratings items, or responding problematically, without being able to directly observe the event, hence their disagreement with the process.

How can institutions and instructors move away from the vulnerabilities of such standardized instruments and toward more direct characterizations of teaching effectiveness? I have argued elsewhere that ratings instruments must first be drawn from the objectives reflecting the intentions of the course, and the events and activities by which the intentions are carried out (Winskowski, 2005). Instructors who take seriously the investigation of their own teaching effectiveness

can design an instrument in a three-step process which ask students to identify 4-8 specific course objectives or meta-objectives, identify the course activities and events that achieve these objectives, and make evaluation items asking students if these course activities/events helped them achieve the relevant objective (Duggan & Winskowski, 2009; Winskowski & Duggan, 2008).

Administrators reluctant to yield the opportunity for comparative evaluation means across classes or departments must at least require that ratings instruments used for summative purposes which address established knowledge on good teaching. The growing field of instructional design theory brings research to bear on the building of a common knowledge base about effective instruction (Reigeluth & Carr-Chellman, 2009). While a detailed account is beyond the scope of this article, an example may be found in Merrill's (2009) "First Principles of Instruction," which distills the following five principles of effective instruction from a number of instructional theories, models, and research efforts: 1. Learning is promoted when learners observe a demonstration. 2. Learning is promoted when learners apply the new knowledge. 3. Learning is promoted when learners engage in a task-centered instructional strategy. 4. Learning is promoted when learners activate relevant prior knowledge or experience. 5. Learning is promoted when learners integrate their new knowledge into their everyday world. Evaluation items which ask students to report on the extent to which principles such as these were implemented and their effectiveness, rather than on superficial classroom events, events beyond students' reasonable knowledge, or mere opinion, will provide a more informative portrayal of an instructor's effectiveness.

In a different approach, Pan, Tan, Ragupathi, Booluck, Roop, and Ip (2009) content-analyzed the written responses of more than 1000 students to questions asking the teachers' strengths and improvements to be made by teachers, to extract key descriptors, both positive (e.g. interesting, approachable, clarity) and negative (e.g., ineffective

lecturing, unclear, ineffective notes). The positive and negative descriptors associated with faculty in the top quintile of student evaluations, based on ratings and written comments, formed a profile of an "effective teacher"; the positive and negative descriptors associated with faculty in the bottom quintile formed a profile of an "ineffective teacher" at their institution. Individual faculty profiles could then be compared to the effective teacher profile. While this approach does not address course objectives or learning models, it could yield well-differentiated profiles of faculty member strengths and weaknesses, as well as hints on how an instructor can improve.

The results of this study illustrates a valuable validity check on the design and construction of evaluation items and have led to a second pilot study, motivated by the students' questionable or unexpected item interpretations and reasons for rating selection. To explore how the data appeared other instructors, a small number of colleagues were invited to assess whether students' responses appeared to be relevant to the item, irrelevant, or not clearly either. The results of this second study will be forthcoming.

* * *

Christine Winskowski *holds a Ph.D. in psychology and a second M.A. 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.*

Author's Note. This paper is based in part on one given with Susan Duggan at JALT 2006 Annual Conference, Kitakyushu. I am indebted to Ms. Duggan for her work orienting and conducting interviews with students in Japanese, for transcribing the interview data and translating it to English, and for editorial comments on this paper. I am additionally indebted to Bern Mulvey for helpful comments and advice.

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. J*ournal of Educational Psychology, 82*(2), 219-231.

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

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52*(3), 446-464.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(4),153-166.

Armstrong, J. S. (1998). Are student ratings of instruction useful? *American Psychologist, 53*, 1223-1224.

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, Fall). A survey of faculty opinions concerning student evaluations of teaching. *California State University, Fullerton Senate Newsletter,* Vol. XIV, No. 1. Retrieved August 29, 2004, from http://faculty/fullerton.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. (1995, September). IDEA Paper No. 32, Student ratings of teaching: The research revisited. Retrieved February 11, 2009, from The IDEA Center http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf
Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco: Jossey-Bass.
Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, (28)*2, 149-160.
Cruse, D. B. (1987). Student evaluations and the university professor: Caveat professor. *Higher Education, 16*, 723-737.
Dickey, D., & Pearson, C. (2005). Recency effect in college student course evaluations. *Practical Assessment, Research & Evaluation, 10(*6). Retrieved March 18, 2010, from http:/pareonline.net/getvn.asp?v=10&n=6
Duggan, S., & Winskowski, C. (2009, November). Expanding the possibilities for course evaluation. JALT (Japan Association of Language Teachers) Annual Pan-SIG Conference, Tokyo.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.
Felder, R. (1993). What Do They Know Anyway? 2. Making Evaluations Effective. *Chem. Engr. Education, 27*(1), 28-29
Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-143). Dordecht, Netherlands: Springer.
Hadley, G. & Yoshioka Hadley, H. (1996). The culture of learning and the good teacher in Japan: An analysis of student views. *The Language Teacher, 20*(9). Retrieved January 4, 2010, from http://jalt-publications.org/tlt/files/96/sept/index.html
Haskell, R. (1997). Academic freedom, tenure, and student evaluation of faculty. *Education Policy Analysis Archives, 5*,

6. Retrieved February 11, 2009, from http://epaa.asu.edu/ojs/article/view/607

Hoyt, D. P., & Lee, E.-J. (2002). Technical Report No. 13: Disciplinary differences in student ratings. Retrieved February 11, 2009, from The IDEA Center http://idea.newbostoncreative.com/sites/default/files/techreport-13_0.pdf

Johnson, V. E. (2002). Teacher course evaluations and student grades: An academic tango. *Chance, 15*(3), 9-16.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. NY: 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.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets, (Eds.), T*he student ratings debate: Are they valid? How can we best use them?* San Francisco: Jossey-Bass.

Lewis, R. (1998, April 27). Student evaluations: Widespread and controversial. *The Scientist 12*(9),12. Retrieved August 29, 2009, from http://www.the-scientist.com/yr1998/apr/prof_980427.html

Marsh, H. W. (1982) SEEQ: A Reliable, Valid and Useful Instrument for Collecting Students' Evaluations of University Teaching. *British Journal of Educational Psychology, 52*, 77-95.

Marsh, H. W. (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). Dordecht, Netherlands: Springer.

Merrill, M. D. (2009). First principles of instruction. In C. M. Reigeluth & A. A. Carr-Chellman (Eds.), *Instructional-design theories and models: Building a common knowledge base, Vol. III*. NY: Routledge.

MEXT (1998). "A Vision of Universities in the 21st Century and Reform Measures" to be Distinctive Universities in a Competitive Environment (University Council Report). Retrieved December 10, 2009, from http://www.mext.go.jp/english/news/1998/10/981010.htm

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox Lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Office of Educational Assessment (2005). Adjusted medians. Retrieved February 26, 2009, from University of Washington, Office of Educational Assessment Web site: www.washington.edu/oea/services/course_eval/uw_seattle/adjusted_medians.html

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. In K. G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations*. San Francisco: Jossey-Bass.

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

Pan, D., Tan, G. S. H., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. K. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education, 50*, 73-100.

Reigeluth, C. M. & Carr-Chellman, A. A. (Eds.). (2009). *Instructional-design theories and models: Building a common knowledge base, Vol. III*. NY: Routledge.

Ruthven-Stuart, P. (2004). Class evaluations; how can they improve education? Paper presented at the annual meeting of the Japan Association of Language Teachers (JALT), Nara.

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

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation,*

4(7). Retrieved June 7, 2005, from PAREonline.net/getvn.asp?v=4&n=7

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

Stake, J. E. (1997). Response to Haskell: Academic freedom, tenure, and student evaluations of faculty. *Educational Policy Analysis Archives, 5*. Retrieved February 12, 2010, from http://olam.ed.asu.edu/epaa/v5n8.html

Trout, P. A. (2000, September/October). Flunking the test: The dismal record of student evaluations. *The Touchstone, Vol. X*, No. 4. Retrieved August 29, 2004, from http://www.rtis.com/touchstone/sept00/20flunk.htm

Williams, W. M., & Ceci, S. J. (September, 1997). How'm I doing? *Change,* 13-23.

Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer. *American Educational Research Journal, 14*, 449-457.

Winskowski, C. (2005). Documenting instructor-effectiveness, Part 1: Vulnerabilities of conventional student ratings instruments (SRIs). *On CUE, 13*(2), 2-15.

Winskowski, C. & Duggan, S. (2007). New directions in the evaluation of university teaching. Liberal Arts, Center for Liberal Arts Education and Research, Iwate Prefectural University, *1*, 1-19.

Winskowski, C., & Duggan, S. (2008). Student opinions challenge course evaluations. Annual JALT (Japan Association of Language Teachers) Conference, Tokyo.

Appendix: Student Ratings Instrument

番	質問項目 Item	あてはまらない This statement is not appropriate	ややあてはまらない…is mainly not appropriate	どちらかと言えばあてはまらない…is more inappropriate than appropriate	どちらと言えばあてはまる…is more appropriate than inappropriate	や⑥やあてはまる…is mainly appropriate	あてはまる…is appropriate
1	この授業にはもともと強い関心があった．(I have had a strong interest in this course from the beginning).	①	②	③	④	⑤	⑥
2	授業内容がシラバスに沿っていた．(The content of the lessons corresponded to the syllabus.)	①	②	③	④	⑤	⑥
3	教え方は要所をおさえていた．(The way of teaching focused on the important points.)	①	②	③	④	⑤	⑥
4	教材（黒板、視聴覚教材、テキスト、配布資料など）の使い方は適切であった．(The use of teaching materials (blackboard, audio-visual aids, textbook, handouts etc.) was appropriate.)	①	②	③	④	⑤	⑥
5	授業内容は量的に適切であった．(The quantity of the lesson content was appropriate.)	①	②	③	④	⑤	⑥
6	教員の話し方はわかりやすかった．(The teacher's explanations were easy to understand.)	①	②	③	④	⑤	⑥
7	授業内容はわかりやすかった．(The lesson content was easy to understand.)	①	②	③	④	⑤	⑥
8	教員は、発言や質問の機会を設け、適切に対応していた．(The teacher provided opportunities for comments and questions and responded appropriately.)	①	②	③	④	⑤	⑥
9	教員の熱意が感じられた．(I was able to sense the teacher's enthusiasm.)	①	②	③	④	⑤	⑥
10	この授業によって知的刺激を受けた．(I felt intellectually stimulated by this course.)	①	②	③	④	⑤	⑥
11	この授業によって得たものは多かった．(I acquired a lot from this course.)	①	②	③	④	⑤	⑥
12	総合的に考えてこの授業に満足できる．(Considering the course as a whole, I feel satisfied with it.)	①	②	③	④	⑤	⑥