Employing Student Evaluations at Japanese Universities:

"Handle with Care!"

David R. Bogdan and Gregory D. Gray

Prologue

Course and teacher evaluations come in many names and (at least superficially different) guises. The table below provides just a few of the acronyms and names given them along with either the source they were cited from or the institution where they are employed.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICES</td>
<td>Instructor and Course Evaluation System (UCIC)</td>
</tr>
<tr>
<td>SEI</td>
<td>Student Evaluation of Instructors (Kent State)</td>
</tr>
<tr>
<td>SEEQ</td>
<td>The Students' Evaluation of Educational Quality (Lawall)</td>
</tr>
<tr>
<td>SET</td>
<td>Student Evaluations of Teaching (Lawall)</td>
</tr>
<tr>
<td>SETE</td>
<td>Student Evaluations of Teacher Effectiveness (Emery)</td>
</tr>
<tr>
<td>SEF</td>
<td>Student Evaluation of Faculty (Haskell)</td>
</tr>
<tr>
<td>SEFP</td>
<td>Student Evaluations of Faculty Performance (Emery)</td>
</tr>
<tr>
<td>SNEF</td>
<td>Student Generated Numerical Evaluation Form (Barrett)</td>
</tr>
<tr>
<td>SOS</td>
<td>Student Opinion Survey (Zakrajsek)</td>
</tr>
<tr>
<td>SOIS</td>
<td>Student Outcomes Information Services (Wuensch)</td>
</tr>
<tr>
<td>SRIC</td>
<td>Student Reactions to Instruction and Courses (KSU IDEA Center)</td>
</tr>
<tr>
<td>SRTE</td>
<td>Student Rating of Teaching Effectiveness (Dooris)</td>
</tr>
<tr>
<td>TFQ</td>
<td>Teaching Feedback Questionnaire (City University of Hong Kong)</td>
</tr>
</tbody>
</table>
While the names may vary, they normally boil down to one thing: students rating, purportedly, some aspect of the education they are receiving. The evaluations often take the form of a series of closed questions usually eliciting quantitative Likert-scale responses, along with, perhaps, a limited number of open questions that allow students some latitude in giving their judgments. Although there are, as the table above indicates, many acronyms, for consistency and convenience, we will use the SETE (Student Evaluations of Teaching Effectiveness) acronym throughout this paper. While there are some differing shades of meaning in the names used, we feel that the evaluations, when used to improve teaching or even to rate teachers for personnel decisions, are, or at least should be, generally rating teaching effectiveness. Accordingly, the final "E" in "SETE" takes on particular importance.

In this paper, we first give a short description of a run-of-the-mill SETE. Following this, we briefly examine the history of the use of course/instructor evaluations, both in the United States and in Japan. We then delve into the controversy surrounding these evaluations through review and commentary on just a fraction of the voluminous research already conducted in this area and by observations stemming from our own experience. General observations, conclusions, and suggestions bring the paper to a close.

The Typical SETE

Likert-scale items are used to measure respondents' feelings or attitudes. Respondents rate a statement by choosing one point along a continuum which may be numbered or lettered, allowing a total numerical value to be calculated from all the responses. The points along the continuum need to be equally-spaced. A scale of "1" to "5" forms a common pattern, and some actually define the Likert scale as being a five-point scale: "1 2 3 4 5".
However, there are those who contend that having an odd number of points may encourage fence-sitting respondents to choose the middle ground, and they propose using a forced-choice response scale with an even number of responses, such as “1 2 3 4”, which would require middle-ground loving respondents to choose either “2” or “3”. Unfortunately, this denies a choice reflecting genuine ambivalence to those who actually do feel neutral about the question. A possible alternative would be to have both an even number of points on the scale along with a separate, additional choice allowing a respondent to express their neutrality. Detailed coverage of the Likert-scale can be found throughout the literature on evaluations. (Brown: 2000, and Trochim: 2000)

Regarding the actual questions or statements used to elicit Likert-scale responses, the preparer of the form needs to avoid ambiguous, leading, or negative statements. Also, as with any test, running a pilot study on a group whose attitudes are known would attest to the validity of the survey, but time constraints in a class often render such an option unlikely.

Open-ended questions, such as “Name three things you liked about the course (or instructor)”, “What three things do you think would improve the course?” or even just “Comments”, allow the students to provide valuable input which can aid the instructor in making improvements.

**Historical Backdrop and Present State of Affairs**

**In the Beginning:**

Although sources (Chang: 2002) indicate that several universities in the US instituted practices for evaluating teaching as far back as 1927, few universities did so throughout the first half of the twentieth century. It was not until the 1960’s, in the milieu of student unrest and calls for greater accountability on the part of the establishment, that student-generated evaluation began in earnest. The early forms
often entailed informal surveys whose results provided students with information about course requirements and grading tendencies of instructors. In its infancy, evaluation of teaching was “infrequent, informal, and unthreatening.” (Trout: 2000)

It was during this heady period that the baby boom had resulted in a substantial increase in the college-aged population, which in turn stimulated an increased demand for higher education. As a result, the number of faculty positions also increased. University administrators slowly implemented procedures to evaluate teaching in the 70’s (Gray and Bermann: 2003). And in the 80’s, legislative bodies controlling government funds for higher education began to apply concerted pressure on institutions to provide more accountability (Ball: 1993) leading to the situation in the 90’s when, as of 1993, nearly 90% of US institutions required some sort of evaluation of teaching (Trout: 2000). That was 10 years ago, but use of SETE shows no signs of being on the wane.

*The Situation in Japan*

Ryan (1998) provides a brief but informative outline of the historical backdrop to the situation with course/instructor evaluations in Japan. Ever since the Ministry of Education introduced its “Guidelines for Establishing New Universities” in 1991 calling for “Self-Check and Evaluation” at universities (Monbusho), there has been an increasing call for evaluation at Japanese universities. Additional pressure for this trend came again in 1996 when the University Council recommended “that tenure be replaced by a contract system for all university teachers, and that one of the factors involved in decisions about contracts should be evaluations of teachers carried out by their students at the end of each semester” (see Ryan: 1998)

The data in the following table (with some adaptations here) mainly compiled from two websites bear witness to the trend toward Japanese universities adopting SETE as a form of self-evaluation.
Table 2: Japanese Universities Using SETE

<table>
<thead>
<tr>
<th>Year</th>
<th>Public</th>
<th>Private</th>
<th>Total</th>
<th>Percent</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>88</td>
<td>184</td>
<td>272</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>197</td>
<td>227</td>
<td>334</td>
<td>55%</td>
<td>607</td>
</tr>
<tr>
<td>1999</td>
<td>119</td>
<td>266</td>
<td>385</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>129</td>
<td>322</td>
<td>451</td>
<td>69%</td>
<td>653</td>
</tr>
<tr>
<td>2001</td>
<td>142</td>
<td>371</td>
<td>513</td>
<td>76%</td>
<td>675</td>
</tr>
</tbody>
</table>

Overall, the figures indicate a clearly defined increase in the number of universities employing SETE over the period covered. The raw numbers show a greater increase than the percentages (where given), signifying that the total number of universities—with those being newly established outnumbering those lost through attrition—has increased over the years covered. Even though the percentage rose at a slower rate than the raw numbers of institutions, it still rose, and at a considerable rate.

Some mention must be made here of where the numbers come from. The total number of universities for the respective years were not explicitly given and had to be estimated using the percentages. The percentages for years 2000 and 2001 were given on the web pages for the Ministry of Education, Culture, Sports, Science and Technology (Monkasho), while the percentage for 1998 came from another source (Campus EOS).

Formative and Summative Use of SETE

Debate and research on SETE have identified two major applications for the evaluation results: formative and summative. The former takes place when the faculty member being rated uses the ratings as a diagnostic tool to identify areas in their teaching methodology that require improvement. The feedback in the form of the data provided by the SETE forms a basis for making developmental decisions.
which aid in the instructor's professional development; in other words, it helps them become a better teacher.

The summative function is seen when ratings by students are used in administrative decisions regarding promotion, retention, tenure, and who receives merit pay or teaching awards, thereby impacting career progress, professional growth, and faculty rewards. Naturally, it is crucial that student rating forms be designed and constructed in such a way as to provide valid and reliable information for these purposes. The recommendation from the University Council in Japan mentioned above constitutes a clear example of summative use.

The summative use generates the most polemics about the validity and applicability of SETE, which is not surprising, as there is money involved. One can see a clear-cut example of these dialectics in the interplay between Paul Trout and Richard Barrett, both professors at Montana universities.

Trout (1997), who teaches English, wrote an article in the fall of 1997, lamenting the change from formative to summative SETE. He indicates that the original goal in the 60's was to use SETE to improve teaching, but that this noble goal has undergone bureaucratization to the point where evaluations are routinely employed in administrative decision-making. He goes on to argue that when evaluations are used to measure consumer-oriented customer (=student) satisfaction, their validity in measuring teaching quality comes into question, because student satisfaction does not necessarily equal student learning. This argument, of course, is contingent on the premise that student learning is an indicator of good teaching.

The following year, Barrett (1998), an economics professor at another university in the state, counters by claiming that Trout misinterpreted some of the evidence and also missed some important implications. He questions the link between evaluations and a watering-down of education, and points out that some of methods given by Trout for obtaining higher ratings will actually improve teaching
quality as well.

Trout (1998) then responds that same year to Barrett’s criticism. He contends that Barrett actually supports his (Trout’s) questioning of the validity of SETE by his admission that, after thousands of studies over 50 years of research, there is still “no consensus” among the experts.

The following data underscore just how important a summative role SETE now play in the US. According to Crombley (1997), the percentage of liberal arts colleges using SETE for personnel decisions jumped from about 67% in 1984 to 86% in 1993, a period of less than a decade. He goes on further to add that “95% of the deans at 220 accredited undergraduate schools ‘always use them as a source of information,’ but only 67% of the department heads relied upon them.”

Let us consider a concrete example of how academic institutions define the dual use of SETE. The Center for Teaching Excellence (CTE) the University of Illinois at Urbana-Champaign (UIUC) identifies two broad areas in which the ratings can be used: for improving teaching and for making administrative decisions. For the former, they describe how an instructor can view the student responses to concrete questions, discuss the suggestions with colleagues who “teach well”, and then make appropriate changes in their own teaching methodology based on the feedback. Faculty members can also pair off in department seminars and discuss their respective ratings. And finally, department heads can use the ratings to identify faculty who may need assistance or guidance with respect to their teaching.

The second category, making administrative decisions, is where the formative use of SETE comes in. The ratings are used in producing a “longitudinal profile” for faculty. These profiles then play a role in promotion and tenure decisions and in for determining the recipients of some teaching awards.

In other words, the data gathered are used to make administrative decisions relating to how the university can make the most of an instructor’s strengths and
reduce their weaknesses. For example, they can be used in determining whether or not to hire a part-time instructor for a full-time position.

**Arguments for the Use of SETE**

Scriven provides a number of compelling arguments for the use of student ratings in comparison with other forms of evaluation for judging teaching effectiveness. First of all, he alludes to the claim by numerous researchers that there is a significant statistical correlation between SETE results and learning gains. [Of course, others have pointed out cases where the correlation is not quite so clear-cut.]

He then goes on to point out the unique position that students occupy in rating their own knowledge gains and changes in their own motivation and attitude towards learning. They are also in a unique position to evaluate directly observable classroom behavior on the part of the instructor, including punctuality, blackboard (or whiteboard, as the case may be) legibility, general enthusiasm, receptiveness to student feedback during the course, relevancy of the tests to the material covered, and other factors.

Finally, keeping in line with a common theme among advocates of students evaluating teachers, he suggests that SETE are the “best available alternative.” (Scriven: 1995) Peer evaluations can be cumbersome and uneconomical. Tracking down graduates for alumni surveys may also pose considerable logistical problems.

While SETE may be the best alternative in evaluation, Scriven also cautions that they should not be the only source of data for the evaluation of teaching merit, a sentiment echoed by others (Seldin: 1993). In addition, he also states that careful consideration should be given to the context in which they are given. The following section deals with some of these context-related issues.
Concerns about the Use of SETE

Since their inception, faculty members and researchers have raised questions about the use of SETE, whether they be used formatively or summatively. These concerns can be roughly categorized as follows: questions concerning the reliability and validity of the instruments of measurement (e.g. the forms themselves); concern about whether the students are qualified to make judgments on teaching effectiveness; the possible effect of contextual factors such as the size of the class, whether the class is an elective or required course, the time the class meets, the status of the instructor, and the difficulty of the subject; the possible influence of demographic (of both the instructor and the students) and personal factors such as age, gender, race, and likeability (of the instructor) on the evaluations; and other concerns such as to how (and when) the SETE is conducted or whether past experiences of the students might effect their judgments.

We will address some of these concerns below. In addition, we will discuss specific concerns that might be more applicable to the situation in Japan.

Reliability and Validity of SETE

Reliability:

In order to be reliable, evaluations must be consistent, stable, and generalizable. Consistency implies that repeated applications should yield the same results. To ensure consistency, the surveys should be administered in a way that filters out conditions that might unduly influence the results. For example, many have emphasized that the instructor should not be present during the evaluations or handle them in any way until the results have been tabulated. Others suggest that they be administered according to a strict time frame.

Stability means that the same respondent would give similar ratings over time. Cashin (1995) cites longitudinal studies that suggest a high correlation between a
student's ratings during the course and those given years later.

Evaluations also need to reflect the general teaching effectiveness of the instructor, irrespective of the particular course they are teaching. Student motivation, discussed below, and other possible extraneous factors can come into play in a cross-course analysis.

*Validity, or “What is being measured?”*:

In order to be valid, the evaluations should measure what they are intended to measure. Logically, this should be the effectiveness of instructor, but even in the bests of circumstances, it is only the students' perception of teaching effectiveness. Quite a few researchers and educators call into question students' ability to rate teaching effectiveness, maintaining that many students may not actually value effectiveness highly, especially if such effectiveness requires more work on their part. Many have termed present-day tertiary education a service industry, and some have pointed out that it occupies a unique position among service industries in that the consumers often want less rather than more. (Crombley: 1997) This is where issues concerning grading leniency and pandering to students come into play.

*Biases*:

Other biases include either factors which have nothing to do with teaching effectiveness or else are beyond the control of the instructor, both of which may affect the results of evaluations. In order for a survey to be valid, such biases must be removed or controlled for.

*Class type*: There is some indication that instructors with discussion classes receive higher overall ratings than other class types and that, within a given class type, smaller classes are generally rated higher than larger ones. Research by Hoyt and Perera (2001) suggest courses emphasizing math/quantitative skill tended to receive lower overall ratings. They found that students in such classes often "were less motivated but worked harder."
Class size: Liaw and Goh (2003), in their study of classes taught at the University of Malaya, found a negative correlation between class size and the ratings the instructors received. In other words, instructors of classes with smaller enrollments tended to receive higher ratings than those of larger classes. One may infer that a smaller class size is more conducive to greater teaching effectiveness by promoting group interaction, better enabling the instructor to play a facilitative role, and allowing the instructor to command greater attention from the students. Students rated the instructors in the following areas: organization, knowledge of subject area, ability to present the material in an interesting manner, clarity of delivery, and appropriateness of examples. Interestingly enough, the results did not suggest a bias due to instructor characteristics such as experience, rank or gender, or other course characteristics such as type and level of subject, or the time and day a course is taught.

Student motivation: Many researchers and academicians have raised concerns about whether a student’s attitude toward a course going into that course may color their perceptions of how the course was taught. Influences related to student motivation include whether a student has chosen to take the class or not, whether the student has had a prior interest in the subject, and what the level of the course is and what field it is in. A report from the Office of Institutional Research at Suffolk County Community College (Suffolk) addresses a number of these concerns. A very short review follows.

Elective vs. required: Common wisdom has it that if a student wants to take a course, as opposed to being forced to take it, they might rate it and the instructor higher, a fact that might lead to instructors teaching elective courses receiving higher ratings than those who teach required courses.

Prior interest in material: Regardless of whether the course is an elective or not, if a student has a prior interest in the subject, they may give higher ratings to the
course. Cashin (1988) points to research showing that students who have prior interest in the subject are more apt to give higher ratings.

*Academic field and level of the course*: Some evidence indicates that instructors of humanities courses tend to receive higher ratings that those of science courses. (Cashin: 1988) Also higher-level courses, especially graduate courses, tend to be rated higher than lower level ones.

One point worth mentioning is that some do not see class size or student motivation as "biases", because students in smaller classes or those with more motivation may actually learn more. Therefore, the teaching is effective, although this effectiveness has little to do with the instructor.

*Grade Inflation Issues*

Controversy surrounds the issue of whether SETE contribute to, or are at least linked in some way, with grade inflation or "pander pollution" (Crombley). UIUC (Date unknown) labels the concern about SETE promoting grade inflation a myth, contending that "easy" graders will not automatically receive higher ratings on evaluations. However, in the same paragraph they point out differing results from research into the correlation between students’ expected grades and how they rate instruction. Furthermore, in their conclusion, they admit there being a complex relationship between grades and rating.

Anthony Greenwald (1997) has conducted studies that suggest a substantial positive correlation between student course evaluations and grading leniency. After noting variations in evaluations of a course he himself taught, Greenwald decided to conduct a large-scale investigation of students at the University of Washington. The results provide convincing evidence that expected grades along with other factors can impact SETE. (Greenwald: 1997, OSU Faculty Club: 1997) Employing multiple analyses, he found a positive relation between expected final course grades (based in part on mid-term grades) and student course evaluations. The
relationship boils down to this: the higher the expected course grade, the higher the evaluation. He uncovered an even stronger relationship when he examined the expected course grade in connection with the student's overall GPA. The analyses indicated that the relation between expected grades and student evaluations was higher when the expected grade stood to increase GPA than when it did not. He also determined that course difficulty is related to ratings. Courses that are more difficult tend to generate lower evaluations.

Greenwald is actually rather guarded in his appraisal of SETE. He does not advocate abandoning them altogether, but rather urges caution in their use including some sort of control for GPA.

**Japan-specific Concerns**

*Students' expectations*: Assuming that students will evaluate faculty members based on some ideal they have for what constitutes a good teacher, we briefly examine three studies into these expectations.

Hadley and Hadley (1996) surveyed students at two Japanese universities by presenting them with the question, “What is a good teacher?” The five most frequent responses, in order of frequency, were “kind”, “friendly”, “impartial”, “understandable”, and “cheerful”. Interestingly enough, attributes such as “knowledgeable”, “intelligent”, and “good teaching methods” occupied considerably lower positions on the list.

Ryan (1998) compares expectations of Japanese students with those of Russian students by subjecting them to an identical question: “What, in your opinion, makes a good foreign language teacher at a university?” The three most frequent responses for the Japanese were: “easy to understand”, “enjoyable lessons”, and “good communication”, as opposed to the Russian students, who placed the most emphasis on “knowledge of subject-matter”, “being demanding”, and “professionalism”.
Although her paper was actually published earlier than that of the Hadleys, Shimizu (1995) had included a twist to a survey conducted in a similar vein by testing whether student expectations differ according to the nationality of the instructor. The results suggested differences in what students perceived were desirable qualities in a foreign teacher and what they wanted in a Japanese faculty member.

*Nationality of the instructor*: Shimizu points out a difference in attitude among Japanese student towards foreign and Japanese instructors. She conducted surveys rating students’ overall impression of classes and on how important they found certain teaching qualities. The results imply that Japanese students evaluate Japanese and foreign instructors by different standards, pointing to a bias in attitudes.

The results imply that foreigners are not seen as serious teachers. Qualities such as intelligence and being knowledgeable were not judged to be as important for foreign teachers as they were for Japanese. On the other hand, students placed more value on qualities such as being easy to get acquainted with and being entertaining for the foreign instructors.

The question one must ask here is whether students’ attitudes towards the nationality of an instructor and their motivation for taking a class with that instructor will effect their evaluation of both the class and the instructor. If so, then a different yardstick should be used when comparing foreign and Japanese faculty; otherwise, one is faced with an “apples and oranges” dilemma.

*Time of class meeting*: At universities in the US, two classes with the same course name and number and receiving the same number of credits will typically meet the same amount of time per semester. Classes normally meet more than one day a week—often in five-day, three-day, or two-day patterns—which allow adjustment for holidays.
In Japan, however, the oft-met pattern is for a 90-minute class to be held one day a week. Unfortunately (or fortunately, depending on your perspective), a majority of Japanese holidays tend fall on Fridays or, perhaps even to a greater degree, on Mondays. At national universities, classes are to meet 15 times per semester, regardless of the day of the week. Make-up days are provided for the contingency of a class falling on a day upon which fall a disproportionate number of holidays that semester.

Contrast this to some private universities, where the situation can be a bit more lax. Let us suppose an instructor at such a private university is teaching two different classes with the same course name and receiving the same amount of credit, but meeting on two different days, with the class on one day meeting more often than that of the other day.

Now let us postulate how such a situation might influence responses to a SETE. Obviously, if an instructor attempts to cover the same amount of material in a different amount of time, questions concerning the pace of the course and the workload should naturally be effected. Workload should also correlate to some degree with the perceived difficulty of the course. Although Cashin (1995) cites research implying that American students actually rate courses with more workload higher, the same may not hold true with Japanese students, especially in light of the results of Ryan’s study. Another category in danger of being affected is “organization”. Students might also frown on the organization of a particular course if the instructor is perceived of as being constantly rushed for time during the class.

In looking at the course schedule for the last two academic years (2002 and 2003) at Matsuyama University, we see that three out of the four semesters have some classes meeting 12 times for the semester, while other classes meet 14 times. The remaining semester has a similar situation, but with the numbers being 13 and
15, respectively. With regard to the first situation, with a difference between 12 and 14 meetings, simple arithmetic shows that an instructor teaching a class meeting 14 times has at least 16% more time to cover the same material that an instructor teaching a class meeting only 12 times. This is a considerable difference.

An instructor teaching the same course on both days, accordingly, faces a serious dilemma: should they adjust the content of the course to meet the time frame allowed, or should they be conscientious about the amount of material a course should cover regardless of the time met. If they follow the latter course, they obviously risk lower ratings on SETE which do not make allowances for such a contingency.

Also coming under the heading of class time is the actual time of day a class meets. While some studies in the US have suggested that the time a class meets (Cashin: 1995) does not impact SETE to a significant degree, similar studies in Japan may yield different results. For one thing, many universities have an actual lunch period between classes. Students and teachers often comment on how lethargic the students (and, unfortunately, many instructors) are at the beginning of the period following lunch.

Having a first-period classes may also impact ratings. Both of the authors of this paper have had experience at various tertiary institutions over an extended period of time in both Japan and the US, and both have noted a greater frequency of students arriving late for a class in Japan. This frequency is even further exaggerated for first-period classes, and one of the major justifications that students give for their tardiness is the very fact that the class begins at such and early hour (8:30 AM), and they also often supply this excuse for notably lackluster participation. If enthusiasm, or making a class interesting, does, as some have suggested, indeed play a positive role in an instructor obtaining higher evaluation numbers, then generating enthusiasm among a bunch of semi-somnolent students in the “wee
hours of the morn” can be challenging, to say the least.

**Statistical Biases**

We often allow our perceptions and interpretations to cloud our judgments when we are asked to rate others, and the impressions of what we have experienced do not necessarily correlate with what we have actually experienced. A number of biasing factors can come into play in such evaluations. Biases in reporting include such factors as primacy effect, similar-to-me effect, contrast effect, halo and horn effects, and distribution errors (severity, leniency, and central tendency). (Kent State)

The first three factors are quite commonsensical. The primacy effect is based on the old adage that “first impressions are important.” The first impression a student has of an instructor quite often greatly influences how they rate them on an evaluation even at the end of the course. People tend to be more kindly disposed to people who have something in common with them, so that the similar-to-me effect might cause a student to give higher ratings to an instructor who shares something with them, for example, similar interests. The contrast effect has students comparing faculty with each other rather than by a performance standard.

The halo effect is often mentioned in critiques of SETE, but what is actually being discussed are two very similar, but distinct, effects having opposite outcomes. Both the halo effect and the horn effect occur when one factor strongly affects the overall ratings, but the former acts positively upon the overall ratings, while the latter lowers them. For example, an instructor, while in reality being a so-so or even poor teacher, could still manage high ratings by telling good jokes in class or being an affable person. On the other hand, an excellent instructor might receive low ratings based mainly on perceived flaws in their personal appearance.

When interpreting results of evaluations, one must also be leery of distributional errors, which occur “when raters tend to use only one part of the
rating scale” (Kent State). In our earlier brief description of the Likert-scale, we mention central tendency where the rater tends to make ratings in the middle of the scale. Severity and leniency occur when the rater holds the instructor to high and low standards, respectively.

The Educator as a Friendly Entertainer

Delucchi & Pelowski (2000), use data from 1,145 SETE forms containing student responses in fifty-seven undergraduate courses to determine what role, if any, instructor likeability plays in student perceptions of learning and how it might effect ratings of overall teaching ability. They found that students who give an instructor high marks in likeability tended to also rate that instructor higher in overall teaching ability. On the other hand, high likeability ratings did not correlate with an increase in their perceived learning. The fact that likeability can have a positive correlation with global teacher ratings, while not affecting the students’ perceived learning, lends some support to those decrying rising consumerism in tertiary education. Students, as customers, expect to be entertained rather than educated.

Williams and Ceci (1997) went even one further by artificially enhancing the entertainment factor while keeping other variables constant. Their experiment at Cornell had the same instructor teaching the same course two different semesters. The content of the course—including the text, reserved readings, syllabus, exams, etc.—remained the same. Recordings were made of the lectures to allow independent raters to establish that the same content was covered. The one difference lay in the teaching style. The course where the more enthusiastic presentation style was used received higher ratings for both the course and the instructor for diverse questions, most of which should have had little to do with enthusiasm. For example, although the material covered was as identical as could be achieved, students from the more “enthusiastic” course felt that they had learned
more although results from the identical exams given in each semester showed that, in fact, they had not. Other course and instructor characteristics receiving higher ratings included text quality, fair grading, accessibility of the instructor, and the instructor’s knowledge of the subject. Obviously, something must be acting upon the ratings if the same textbook receives significantly different ratings in the two surveys. We need to look for other influences and provide controls for them.

**General Observations**

Making a class entertaining is not necessarily a bad thing. A student who has enjoyed a course might have more incentive to continue their education in that area. However, entertainment is not the primary goal, and learning must not be sacrificed for the sake of entertainment. That is, of course, if one accepts the premise that the goal of a college student is an education rather than just receiving a diploma.

In addition to concerns about the evaluation forms themselves and how they are administered, one must also ensure that the results are utilized fairly. Both of the authors of this paper are tenured English teachers, so we take this opportunity to examine how SETE might affect the quality of English education in Japan, because we have a vested interest in seeing that anything that influences this quality does so in a positive way.

Foreign faculty members in non-tenured positions at Japanese universities could be especially vulnerable to the whimsies of students. For example, in addition to the three native speakers of English in tenured positions at Matsuyama University (the last position for a tenured native speaker of English was filled over ten years ago), there are four positions for so-called “Special Foreign Instructors”. These are two-year positions with the possibility for a one-year extension. If this extension is based, even in part, on the results of SETE, and the students catch wind of the fact that their ratings can decide whether the foreign instructor gets another year or not,
this could give them certain power over the instructor (at least in the mind of the instructor even if not in actual fact).

Our university also has a very high percentage of part-time instructors teaching English classes on essentially what are one-year, renewable contracts. If the SETE results are used in deciding their retention, the same problem occurs. Some may feel that pandering to student desires may be the way to keep their job, and, if the students’ desires run in the direction of “less work”, English education suffers.

This situation is not limited to Matsuyama University. There are actually universities in Japan that will offer the same English-teaching position to Japanese and non-Japanese (often native English speakers), with the proviso that Japanese will automatically be tenured when they are hired, while the hapless foreigners are relegated to term-contract positions, whose extension may be contingent on, among other things, results of SETE.

In the above circumstances, SETE results will have a different meaning to tenured and adjunct faculty, which might lead to grade inflation or influence teaching in other ways not conducive to quality education. In this special case, English education will suffer, but the same concern applies to other fields.

**Conclusion**

Although this paper might appear to be heavily weighted in the area of concerns about SETE, neither of the authors advocates an abandonment of their use in Japan. What we do urge however is care in their preparation and administration, and extreme caution in their application, especially with regard to summative decision-making. If they are to be employed in the summative sense, then instructors must be rated on a level playing field, if, in fact, fair comparisons among all instructors at an institution can be carried out.

To accomplish the above, any possible biasing factors, such as those mentioned
above, must be investigated beforehand and removed or at least accounted for in the rating system. Questions not directly relating to teaching effectiveness (for example, concerning the teacher's appearance or attire) must either be weeded out or edited, as must those relating to factors beyond the instructor's control. An example of a factor beyond the instructor's control would be to have a student who never once went to the teacher's office answer a question about whether that teacher kept regular office hours.

The Japanese have, in recent times, gotten wise to the possible pitfalls encountered when blindly following the lead of the US. What works in one culture can occasionally go awry in another. While SETE have seen widespread use in the US, the same use might not be appropriate here. And while they have been around for quite a while, and are touted by some as the definitive answer to faculty development, many, even among American researchers and faculty, still question how they are employed.

The jury is still out on whether summative evaluations contribute to grade inflation, but even if it does come out that the influence is negligible or that it can be controlled, if some faculty members believe that pandering to the students will lead to better ratings, grade inflation could still be a problem. Birnbaum (2000) discusses similar concerns.

To sum up, evaluations must be developed, administered, and utilized with accuracy and fairness in mind. This is especially true when the results are to be used in a summative fashion. Keep the playing field level for all players; jobs and salaries are at stake!

Suggestions

It is almost a cliché at this point in a paper to call for future research, and we would be remiss if we did follow in this fine tradition. We have covered a few of
the numerous concerns about SETE which have been voiced in the mass of literature on the subject and have also provided concrete examples of potential concerns based upon our own experience and which may be specific to Japan. Those concerns need to be addressed, and the most appropriate way to address them is to subject them to verifiable testing to determine whether they are indeed warranted.

While future studies on the subject are definitely in order, administrators of SETE can still at present pay heed to time-tried guidelines relating to their administration and use. For example, if used in the formative sense, they could be administered partway into the term to allow the instructor to make the necessary changes for that course.

Cashin (1990) provides a list of clear-cut, easy to understand recommendations for how to use evaluations. Among the points he stresses are: using multiple sources for evaluation, care in administration of student based evaluations, allowing flexibility, deciding on controls for bias, and various suggestions for the forms themselves. In addition to the typical closed Likert-scale questions, open-ended questions provide essential formative input.

Accurate results demand standardized, systematic, and fair administration of the SETE. They should be administered according to a strict time frame for all courses, and instructors should never be present while the students are filling them out. In fact, instructors should not handle the forms at all. A reasonable percentage of the students in a particular course (Seldin suggests a minimum of 75%) should respond to the questionnaires.

Evaluations for just one course or one semester should not be the sole basis for summative decision-making. The administration needs to look at multiple courses over a period of several semesters to make judgments. In fact, much of the literature we researched stresses that SETE should not by themselves be the basis for professional decisions; they should form just one of the criteria.
In addition, institutions employing SETE must have some clear goals in mind other than just regarding the results solely as a “carrot and stick” device. If a faculty member consistently receives lower ratings, then true faculty development demands that the school have some sort of assistance program to help the instructor become a better teacher. Workshops and peer counseling are just two mechanisms which come to mind.

References


Employing Student Evaluations at Japanese Universities: "Handle with Care!"


