# Benefit or Bane?: Teachers' Reflections on the Usefulness of Student Evaluation of Teaching Surveys

Peter Burden

*Okayama Shoka University*

西日本の大学で教鞭を取る常勤または非常勤のEFL講師22名が自発的に1時間のインタビューに答え、自身の大学に於けるSET（学生による授業の評価）アンケート調査の導入について所信を述べた。彼らの回答からは、現在に於ける大学管理運営の環境においてSET調査は有用性を欠くことが示唆された。教師たちは、SETが不適切な時期に実施されるために、この評価ツールのもたらす結果が知らされず、曖昧になってしまうことから、自分が不公平な評価を受けているのではないかと恐れており、またこの評価の目的についても説明を受けていない。本調査に参加した教師たちは、SETの質問内容に対する意見も聞かれぬまま、SET調査を実施することだけを求められている。そして、評価されることへの強迫観念から、調査の水準は低いものとなり、教師がそこから得る知識は、あったとしても非常に少ないものである。また、調査から何らかのフィードバックを得られとしても、その価値を疑問視する者が多い。

## Introduction

It has been suggested that Japan has entered an "epoch-making phase in the history of higher education" (Arimoto, 1997, p. 206) whereby a "Big Bang" (Goodman, 2005, p. 2) has led to a government initiated rush for reform. The introduction of student evaluation of teaching (SET) surveys is one part of sweeping changes in the ways universities are organized in response to market forces.

The Ministry of Education, Culture, Sports, Science and Technology

(MEXT, 2004) has advanced SET surveys as a principal method of getting information from students as a reflection of a more economic-centered, more market-sensitive, decentralization movement which emerged at the start of a new millennium in partial response to the expected decline in the 18 year-old population (Yamada, 2001). As universities' total capacity will equal the number of applicants in 2009 (Tsurata, 2003), universities are now subject to a buyer's market where students are "courted customers rather than supplicants for admission" (Kitamura, 1997, p. 145). While the population is shrinking, the percentage proceeding to some form of post-secondary education has reached 72% (MEXT, 2004, p. 3). As a higher percentage of high school graduates are entering tertiary education, the content of university education must change to meet diverse students' abilities and knowledge with more consideration required to make education more attractive (Yamamoto, 2005). Coupled with this, in a time of economic retrenchment, the public has become more critical of government spending, so a market agenda underpinning government strategy to determine accountability in education has led to the introduction of SET as part of wider evaluation to encourage institutions to take more responsibility for their management. This attempt at accountability suggests a results-oriented milieu in which the government demands that the public investment in education justifies closer scrutiny of the outcomes achieved.

## Rationale for this Study

This study looks at teacher perceptions of the introduction of SET as a follow-up to my previous research (Burden, 2005) into tertiary students' views which found that 76% of 203 questionnaire respondents thought that evaluating through SET surveys was useful. However, only 21% said they were clearly told how evaluations are used and 42% thought the process could be better. Worryingly, 36% of students felt that teachers who receive consistently poor evaluations should lose their job, while 43% felt that such teachers should get a

salary cut. However, many students appear to be going through an end-of-semester ritual having to fill in the same form many times. Evaluation lacks "tangible immediacy" (Spencer & Schmelkin, 2002, p.406) as expressing opinions does not in the long run benefit that particular student, which may lead to perfunctory student response that can affect teacher feedback and also, possibly, their futures. Whether SET surveys provide valid, reliable information depends on the quality of the processes used to obtain information and on the ability of receivers to use it appropriately," note Theall and Franklin (1990, p.19), but previous findings (Centra, 1993) suggest that if teachers do not believe that the evaluation information they receive has value, and the source is respected, they may simply dismiss it. There is little in the literature on the effects of the introduction of SET on EFL professionals in Japan and this interview study sought to gain an understanding of the perspectives of teachers.

## The Study

Previous studies (see Spencer & Schmelkin, 2002) largely utilized questionnaires which "test" how prevalent beliefs are among teachers, and, like SET surveys themselves, are easily scored but have a restricted range of responses. This study, being interpretive, seeks to understand teacher views of the introduction of SET and *why* teachers hold their views. In February 2005, I sent an introductory e-mail to JALT-connected local and expatriate teachers, briefly explaining that I sought teachers' views of their experience of university-driven teacher evaluation. With teachers who expressed an interest and were willing to volunteer time, a semi-structured interview of around an hour was arranged with each participant. Although an emerging design, an interview protocol that identified the following key questions which encouraged follow-up questions was used in interviews. These questions asked:

1) What "benefits" have teachers gained from the introduction of SET surveys?

2) What are EFL teachers' understandings of the purposes of SET?

3) How are the SET surveys administered in their respective universities?

Twenty-two full time tenured local and expatriate teachers; limited (or fixed) term contracted local and expatriate teachers; or part time local and expatriate teachers volunteered to take part. They all worked within tertiary institutions within one city in Western Japan. They ranged in ages from their early thirties to late fifties, and in experience from a single year to over thirty years teaching in the tertiary sector. I was aware that for some of the local teachers being interviewed in their second language might be a little stressful, so I emphasized that there were no "right" or "wrong" answers and that I wished to gain insight into their experiences of evaluation within their institutions. As evaluation is inherently political, I outlined anonymity and confidentiality procedures, stated that I would want to record the interviews, and added that the participants could withdraw at any time. I requested a copy of their university evaluation form at the interview so we could use it as a reference to discuss any questions that arose during the interview.

The participants and I negotiated a convenient time and location for the interview which were often held in the teacher's office or in the part time teachers' rooms where they worked. The interviews took place between March and May 2005, approximately two months after teachers had administered evaluation during the final weeks of the second semester of the previous academic year which ended in late January or early February. I imagined that the university administration would have had sufficient time to analyze and send the data back to teachers in anticipation of the new school year which commences in April. However, none of the participants had received any feedback

at all despite the two month gap and the fact that the new academic year had either commenced or was about to. The participants freely volunteered their time and consented to the interviews being recorded. The transcriptions were returned to the participants for comments and all the participants chose pseudonyms which are used in this study.

## Data Analysis

Through coding, all the comments for each question were ordered into categories using the "constant comparative method" (Lincoln & Guba, 1985). Key word analysis from paragraph to sentence generated a great deal of data as teachers expressed more than one comment on a question. The categories were quantified by identifying themes and isolating data by the number of occurrences (Miles & Huberman, 1994), so I looked at each participant's transcription in terms of units of information that became the basis for defining categories.

## Findings and Discussion

The following section discusses the benefits of SET surveys for individual teachers and their perceptions of the evaluation purpose, and illustrates the administration procedures in the participants' schools. While participants would agree with Norris' (2006) view that evaluation, as part of assessment, should act as "an agent of supportive program enlightenment and change" (p.578), findings suggest that individual teachers understand neither the criteria nor the process of this kind of evaluation. There is a lack of shared understanding of the mechanics of the evaluation system, that the evaluation criteria have a clear, consistent rationale, and that SET surveys represent the most important aspects of teaching.

## The Benefits of SET Surveys

All of the participants accepted that evaluation is necessary as a

process to give insights to teachers: all the teachers wished for an open, improvement-focused, cooperative evaluation but, crucially, one that utilizes more than one data source. The results in Table 1 show the benefits of SET that teachers gained from their use in their respective institutions. Ten teachers believed that good evaluation results give confidence or reinforcement, with Jack, for one, noting:

> I have never had a negative comment about my teaching or myself personally which is good. The best comments, usually about three a class, are, 'I didn't like English at school, but I have enjoyed using it in class.' These comments mean I have achieved my goal.

Gaining high scores may justify teachers' classroom approaches, which does not however encourage change. Teachers felt that they could gain "hints or tips" or that the free comments section had "occasional thought provoking" comments, while eight teachers suggested they could learn if they were above average, and four felt that an evaluation "process" (but not necessarily SET) can make teachers more conscientious. However, for other teachers the comments were seen as being far too general, as Melvyn suggested:

> The positive comments just reinforce what I already thought, the negative ones they hint a little bit, they make me think, but I already felt that with daily interaction with students. I already know when I'm not getting through.

Julian suggested SET surveys reinforced "minimum competence" and Miki thought they might be "effective for teachers who are really bad but for passionate teachers who have their own styles they may have some complaints." As Julian stated:

> Whereas the surveys don't portray the whole picture there's nothing

in them that is not part of good teaching. You have to have some kind of minimum competence at least and there's nothing in there that I wouldn't consider at least in part of a minimally competent teacher.

SET surveys focus almost entirely on identifying poor teaching. In doing so, student evaluation may have lost the power to bring about improvement.

Table 1

*The value of SET surveys as stated by the participants*

| The value of SET surveys | Number of participants mentioning item (N = 22) |
|---|---|
| Comments about benefits | |
| Positive evaluation can give confidence and reinforcement | 10 |
| Can learn if teachers are better than the average | 8 |
| An evaluation process can make teachers more conscientious | 4 |
| Teachers can learn about their teaching approach | 3 |
| Subtotal | 25 |
| Comments about the lack of benefit | |
| Student comments are too general | 11 |
| Concern over the use outweighs any benefits | 8 |
| Lack of any follow up | 5 |
| Results are difficult to digest | 4 |
| Teachers are too afraid to read the results | 2 |
| The use of SET surveys in their present form is completely worthless | 2 |
| Subtotal | 32 |
| Total comments | 57 |

Note: Participants may have made more than one comment in each category.

Reflecting this, eleven teachers said that student comments were far too general. Many teachers observed a lack of thought and perfunctory

feedback invested by students who "kind of shoot" or "whip" through the survey, or just "zap it out," while they noticed from feedback data that scores of all three, all four or even all five were common. Nearly all of the participants referred to the formulaic nature of the responses. Melvin noted that the "biggest response is the lack of response" receiving an average of five forms with comments out of a class of 25 or 30. Keiko added that "most of the students didn't write comments because they didn't have any interest and just wanted to leave as fast as possible." Miki's students complained of tiredness and some said, "Again? I've done this many times." Eight teachers were concerned about the use, with Ken stating that any kind of evaluation can be "nerve-wrecking":

> Maybe I had the feeling or hope or desire that they would give me some useful information that I could work with and I guess that the other thing is, to return to the administration issue, I'm very concerned about how the data is used by the university.

Participants voiced concern over data use as students are empowered to influence careers. Ed, for one, was "wary" of ranking teachers which suggests "winning and losing" which will negatively affect teacher motivation or morale. Five teachers lamented the lack of feedback or follow up with Ken wondering about the amount of time the evaluating body took to "process a pile of machine readable cards." Pat, too, emphasized the perfunctory nature of SET and said that:

> Maybe we shouldn't expect a pat on the back, but it seems we administer the stuff, we collect it in, they process and analyze the data and someone makes a nice little graph and fills out the form and it's left at that.

Because there is so little information of any worth, and little or no

mechanism for remedial help, the potential for teacher growth is limited. Teachers added in the course of interviews that results were too late even to inform the second semester: For Ed it took four months for feedback to come back, for Jane eight, while Robert thought that results were collated years later in "some kind" of book. Even getting data does not always guarantee it is understandable as four teachers suggested. Ken said:

> You get the results on a sheet and your scores are compared to other faculties. I don't know how it works. It's all homogeneous. Where do you stand? I mean, you get a 4.8 and someone else gets a 4.2 and that is 0.6 of a difference. Is that important?

For Steve, interpreting the data could be challenging and difficult to digest for a busy teacher who cannot read formal Japanese so well. Jack and Ken said all the graphs looked the same so they did not pay attention to the data reinforcing the technical, depersonalized top-down nature of evaluation. Ken observed:

> I look at the scores comparing me to the faculty average. And that is it. I look at the scores and hope I am above average but what does that mean? Does this information go anywhere to revitalize the curriculum? I doubt it.

Two teachers saw this form of evaluation as completely meaningless, with Robert suggesting that cross-curricular administration renders the questions so vague as to be "a complete waste of time. Some of the questions have no application to any class I teach." This suggests that the survey's face validity, defined as the extent to which an instrument looks as if it measures what it is intended to measure (Patton, 1996, p.253), is compromised as teachers do not appreciate the implications of comparing different classes.

Teachers also questioned the overall content validity of the surveys. Braskamp, Brandenburg and Ory (1984, p. 20), contend that there must be a link between teacher performance, how evaluation is carried out, and reward for quality of performance in order for any evaluation mechanism to have utility. Credible evaluation means attaching importance to gaining usable data and ensuring that teachers accept student judgments and also, importantly, the manner in which data is used. However, teachers' perceptions of and willingness to accept the use of SET were colored by what they felt was the purpose for its introduction and the effects it could have on status and continued employment. Miki and Ayumi, who are both single, part time teachers, were "too afraid" to look at the comments: Miki "left them in a drawer," while Ayumi just felt enormous relief to get above average results. This consequence of evaluation will be referred to again in the next section.

## Teacher Perceptions of the Evaluation Purpose

Table 2 summarizes participants' perceptions of the purpose of SET surveys, with most participants suggesting more than one purpose. Ambiguity over purpose is obvious as 15 participants thought that while they considered evaluation *may* be used for teacher retention decisions, 12 thought it was just a "ritual" that administrators and teachers engage in because it is expected - not because it is valued. Participants also voiced concerns over the quality and timing of the SET and 7 teachers said that they were unaware of an explicit purpose, but suggested that implicitly SET is used for "assessment," "as a way of "watching over" or to "get a detailed view" of teachers' practice. For many, the timing was seen to influence the purpose while the lack of feedback data compounded participants' questioning of the evaluation purpose.

Table 2

Participant statements beliefs of the purpose(s) of SET surveys

| Reasons/Purposes | Number of participants mentioning item (N = 22) |
|---|---|
| For teacher retention decisions | 15 |
| University ritual | 12 |
| Used for assessing teachers | 7 |
| To conform with MEXT guidelines | 4 |
| To emulate America or the West | 4 |
| For teacher feedback | 4 |
| As a way of gaining student opinions | 4 |
| Do not know | 8 |
| Total | 58 |

Therefore, unlike some earlier studies (for example, Schmelkin, Spencer, & Gellman, 1997), many participants found the link between teacher evaluation and actual course improvement at best tenuous. Eleanor remarked that:

> It has never been made clear to me how the evaluations are used, who sees them, how the information is stored, who has access to them and for how long they are stored. The confidentiality and access issues are similarly important. It has never been made clear to me whether the focus of the evaluation is the teacher or the course.

Stephanie, a part time teacher at both a private and a national university, was quite clear in her own mind that there was only "one reason" for the administration requiring the information and that was "so they can get rid of under-performing teachers." She couldn't see any evidence either in the questions, or in the way the evaluations were administered, that indicated that the purpose was for a teacher's own personal development. Stephanie concluded that, "It never entered my mind that some people might consider that to be the purpose." Taut and Brauns (2003) suggest that implicit power differences and mismatches between teachers' and administrators' concept of social

reality requires conflict resolution because of a loss of teacher control which leads to insecurity concerning consequences of the evaluation and thus removes any benefit to act as a tool for improvement. This lack of knowledge over its intended use lead to feelings of suspicion so that participants saw the administrative body as "shadowy" or "unknown," and used metaphors of a "guillotine" or "weapon" to describe the potential consequences of a poor evaluation.

Four teachers saw universities as simply conforming to guidelines due to the lack of articulation of policy: "Keeping the *Monbukagakusho* [Ministry of Education] happy" Steve suggested as the main purpose, and added that "if you're looking at the ritual side of things, the universities need to be seen to be doing something if they want to get *Monbusho* money. I can see that being the major motivation for them to do it." Purpose for cash-strapped schools is linked to grants and subsidies which are performed because it is expected as part of the atmosphere or trend in society. Again, this also reinforces the lack of clarity of policy decisions imposed through central government imposed dictates. Four teachers suggested that SET is to emulate America or "The West" and surmised that for MEXT, American universities are seen as "advanced," and evaluation is an imitation of American policy based on initially superficial understandings. Derek was similarly skeptical about insufficiently thought-out implementation noting:

> Particularly in this country people are fashion-conscious and copy what someone else is doing and introduce things. But once it's introduced it's very difficult to stop so the thing goes round and round in automatic fashion regardless of the meaning of it. If someone thinks it is a good idea it's adopted regardless of whether it is a good idea or not.

Yet, *if* SET surveys are for formative purposes, that is, for improvement or to make better teachers, participants lamented the lack of clear

directives as to *how* evaluation for diagnostic improvement should be encouraged, beyond prescribing the *need* for improvement against an overall mean score. Only 4 of the 22 participants saw evaluation as providing "teacher feedback" or for a way of getting students opinions'. The feedback was seen as vague and of poor quality as noted earlier. Jack thought that:

> If the universities did use the information for diagnostic purposes to get the teachers together and say, 'perhaps we're doing this wrong' that would be really good, but I don't believe they do.

While evaluation has been seemingly accepted by psychometric researchers as rigorous and technically valid in the field away from the chalkface, how they are used in day to day practice remains problematic. SETs are "misinterpreted, misused and not accompanied by other information that allows users to make sound decisions" say Theall and Franklin (2001, p. 47), which leads to poor decision making by the administrative body that acts on their results in summative practice. Lack of knowledge for teachers about the use creates "a smoldering dynamite fuse" (Patton, 1996, p. 22) of suspicion that may not even be warranted. Negative perceptions including denial and defensiveness, and, conversely, feelings of elation and relief, can interfere with the information itself and thus have a bearing on whether this information will give rise to any active professional development (see Moore & Kuol, 2005). Jane, a contracted teacher, pointed to the short term, revolving door of contract negotiations. This gave her a lot of stress around contract-renewal time so "If they [administrators] need to shed staff they'll find a way to interpret these results so they can." She found the lack of criteria for how SET surveys are used or how the results are evaluated, and saw the failure of the administrators to transmit in detail the intended criteria, as "extremely problematic." In particular, it leads to teachers ignoring data potentially useful for faculty development

because they mistrust the data sources. In this study, teachers noted that evaluation has been very recently introduced in their respective institutions, so the administrative body seemes not to have formulated a clear structured policy of its use, heightening teacher fears and cynicism over its purpose.

## Evaluation Purpose as Dictated by the Administration Procedure

Ideally, if evaluation is for formative purposes, appraisal should take place around mid-term to initiate actions in response to student concerns within the lifetime of the course. As Alderson (1992) notes, if evaluation is left to the end of a course it loses any opportunity to inform and influence teaching. Students need to be part of the feedback loop by seeing the results of their action being implemented during the lifetime of the course (Ballantyne, 1998), and Expectancy theory of motivation (Chen & Hoshower, 2003) suggests the process is only "worthwhile" to students when they realize the personal benefits of investing their input (Dunegan & Hrivnak, 2003).

Table 3 shows the timing as observed by the participants in their schools. SET surveys were often handed out on the day of the final test. Many teachers noted that the final test occupied only a part of the final class and that evaluation forms were given out after the test has ended. Steve recalled that "All they did was give me an envelope with the things inside all in Japanese. They asked me to work it out for myself." With little or no advice, administering SET in the final class can distort results (Aleamoni, 1981), communicate that it is an afterthought, somehow unimportant, or as a final judgment on performance (Dunegan & Hrivnak, 2003), or an unwanted distraction before the test (Cashin, 1990). Often teachers felt pressured for time allocating just five minutes. So with such little time to complete, only "the most frustrated…student will provide more than cursory input"

(Dunegan & Hrivnak, 2003, p. 296). Ayumi suggested that the delay in getting feedback impacts on teachers' abilities to use the results for improvement:

> I think this is something that I want the school to improve because the first semester evaluation was given to me in the middle of the second semester so I can't really change in my first 4 or 5 classes. A good time would be two weeks before the class starts, ideally.

Table 3

*The timing of the SET surveys*

| When teachers are asked to administer the surveys | Number of participants mentioning item (N = 22) |
|---|---|
| In the final few weeks of the semester | 6 |
| In the final class | 4 |
| In the week before the final test | 3 |
| Within a two week period before the final test | 3 |
| Teachers are just given a date by which to do it | 2 |
| Teachers just find the forms in their mail boxes | 2 |
| No data | 2 |
| Total | 22 |

Linked with timing, teachers' practices during administration inadvertently influence SET scores. Literature suggests that student ratings are higher if the instructor remains in the room during administration, give out and collect forms themselves (Wachtel, 1998), or when students are told that ratings will be used for personnel decisions (Seldin, 1993; Cashin, 1995). Teachers in this study did not know the official procedure: Melvin thought "I'm supposed to disappear," and other participants recalled office staff entering the classroom and collecting the papers either for "objectivity" or due to a lack of trust that "teaching staff are not hiding in the corner and erasing answers and filling them in for their own benefit." Arguably, if evaluation is for developmental purposes, there should not be a need for rigorous supervision as the aim is to provide information for individual teacher

improvement.

While the majority of teachers were expected to give the evaluation forms in all their classes which may number up to fourteen a week, at least two institutions allowed teachers to choose one or two classes. As it is questionable whether ratings objectively reflect the qualities they are designed to measure, and capture factors normatively irrelevant to teaching quality, so teachers' implementation of evaluation at different times can change ratings scores. Some teachers stated that they chose the classes with the most students so they could get potentially more feedback from students, but as Satoko noted, "specialist" teachers who teach seminar classes and wish to get "high scores" can choose classes which often consist of only a few students and of higher course levels both of which affects reliability (Braskamp et al., 1984; Cashin, 1990).

## How to Make Evaluation Beneficial

"The paradox of teacher evaluation is that it holds the potential to help nearly every teacher improve, yet in actual practice it helps almost no one" state Stiggins and Duke, (1988, p.1). Although twenty years have passed, the paradox remains that SET surveys seem to have done little for teacher improvement (a "benefit"), with negative consequences (a "bane") foremost in the minds of many of the participants in this study. From a teacher perspective, the link between teacher "evaluation" and actual course "improvement" through SET survey use is often seen as tenuous. With no explanation from the administration, the "purpose" becomes even more unclear, heightening tension, reducing autonomy and freedom, and contradicting the notion of an effective institution as one that "simultaneously fosters individual faculty development and fulfills its own institutional goals" (Braskamp & Ory, 1994, p.20). Boise suggests (1992, p. 248) that a lack of explanation leads to a "vagueness" in qualities being rated, so that teachers often react to

them defensively, while "reactance" (Taut & Brauns, 2003, p.252) can occur through perceived restrictions on teacher behavior and where the consequences are beyond their influence, or unpredictable. A lack of clarity of the evaluation purpose may also encourage academically irrelevant practices to raise SET scores including making courses easier, lowering standards, offering inducements such as saying that future income would be affected, and administering SET surveys when poor students skip class (see for example, Simpson & Siguaw, 2000).

If SET surveys are for diagnostic improvement, a more useful process would be for teachers to co-create evaluations to tailor questions to reflect their own classrooms, and to administer them during the semester at mid-term so that teachers can make use of feedback and students can see change in the lifetime of the class. At present, those who make survey questions are nameless, often demonized, persons such as "big cheeses," "old farts," and "groups of little men" who form "nameless committees" and "get together" in "darkened rooms." The participants had little confidence in the ability of power holders whose views are not consonant with teachers' educational goals and conceptions of teaching. Participants suggested that questions were outdated and irrelevant being written "about a million years ago" by some "Japanese statistician type" or by "someone in the hard sciences a long time ago." A technical, rationalist view underpins question choice, and teachers administer and receive the same questions each time which loses the power to inform and make change (Centra, 1993). Therefore, many participants suggested the need for more teacher involvement, more dialogue between teachers in order to discuss the results to aid the reflective process for change, and to remove competitive feelings and the pervasive atmosphere of secrecy that surrounds data results. Any form of threat leads to denial which will interfere with information and influence any active teacher professional development. As Johnson (2000, p. 431) eloquently suggests, teachers feel a "catharsis of relief at not being revealed as incompetent" which reinforces current teaching

practice and deters faculty from exploring and using a variety of teaching methods.

Stephanie, an informant in this study, suggested that "line manager" meetings in which teachers are comfortable to talk about problems (and successes) should be set up. Similarly, Melvin argued that "issues and complaints" needed to be brought up "person to person" where there is some comparison and future plans formulated. Jack and Pat thought a forum would encourage teachers to "come up with a game plan" based on questionnaire results. In order for teachers to focus on improvement and not *just* as an implementation of bureaucratic dictates, at first there must be an explanation of purpose to allay feelings of threat.

Essential for improvement is an environment which:

1) "Dignifies" evaluation
2) Has a greater role for students
3) Encourages a pragmatic stance by teachers and recognizes that evaluation is here to stay
4) Recognizes the need for horizontal dialogue

## Dignifying Evaluation

Marincovich (1999) writing as a university administrator argues for a concerted effort on the part of administrators to "dignify" (p.52) teaching evaluation. She suggests there is a need to make its implementation more efficient and appealing which would influence faculty members' receptivity. Stephanie suggested "if something is going to be useful it should be discussed," echoing Canagarajah (2006) who argues for a "ground-up construction" (p.29) to encourage a sense of participation in evaluation. There is a need for more than one evaluation method, especially the introduction of peer observation in recognition of the multi-dimensionality of teaching; one set of

criteria cannot fit all teaching situations, and that having only one tool undervalues teaching.

## A Greater Role for Students

Validity is compromised through learner indifference. Many participants were unable to adequately explain the evaluation rationale to the students which then may have influenced their SET scores. Students also need a greater sense of involvement by seeing their contribution to evaluation acknowledged. Students may have a cynical attitude which teachers themselves may promote through their own indifferent administration in the face of doubts about the *value* of student data. As Spencer and Schmelkin (2002) note, "Since students are unsure whether their opinions matter, or to what purpose the ratings are put, they may not pay attention to them" (p. 406), in contrast to the stated desire to provide feedback, so there needs to be a change in the school ethos to one where opinions are freely exchanged without threat or feelings of comparisons. Perhaps students could be encouraged to keep a diary so they are "sensitized to [the] recency effect" (Dickey & Pearson, 2005, p.8), where students only evaluate the most recent class, to promote a source of stable course evaluations. In this vein, Pat said that he gives students an overview of the semester "so students can link what they've done with what they've achieved" because without "refreshing their memories" the evaluation "wouldn't reflect what they've done." John likewise suggested that reminding the students what they did over the whole course including assignments and in-class activities gives students a sense of progression as "recency" is also an "understandable error that is based on the difficulty to remember performance" (Dickey & Pearson, 2005, p. 9) in a semester that may be close to four months old.

# A Pragmatic Stance: Recognizing That Evaluation is Here to Stay

Many participants in this research administered their own evaluation alongside the "official" mechanism. One can maybe change the system by energetically showing how *their* evaluation "works." If teachers want to change the evaluation ethos to one of improvement and not *just* an implementation of bureaucratic dictates, one could show the results to the "powers that be," and outline *why* letting each department create their own questions for the evaluations to remove the broad, generic, "one size fits all" questions that appear on most evaluations is important. However, this requires offering to work with the office staff to rewrite school-wide evaluation recognize that few people wish to do extra work, but demonstrating a new idea and showing willingness to do the ground work heightens receptivity to the idea. Any attempts to start or change a university evaluation system should always emphasize mutual benefits: "Working the system" by quietly talking to as many people as possible, including the few people who actually make decisions, is crucial.

## The Need for Horizontal Dialogue

An important step in improvement is knowing how to establish dialogue. Effective schools are characterized by administrators who are open to change, are flexible, and who encourage teacher innovation (Marincovich, 1999). Evaluation should be carried out in a "context of trust and development rather than fear and censure" (Moore & Kuol, 2005, p. 60). Mutual trust and respect must be encouraged to show confidence in the individual and collective competencies of department members, and a convergence between administrators and teachers in accepting both the goals and the means of evaluation. Through dialogue, we can agree on the steps to good practice (Theall & Franklin, 2001) such as what will be evaluated, who will

contribute and what criteria will be used while including resources for improvement and support of teaching and teachers. Without a fear of retaliation, faculty need to air differences of opinion constructively without imposing a single viewpoint and including part-time and "limited term" contracted teachers' perspectives, who otherwise will perceive their role in carrying out work productively as unimportant.

## References

Alderson, J. (1992). Guidelines for the evaluation of language education. In J. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 274-304). Cambridge: Cambridge University Press.

Aleamoni, L. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp.110-141). Beverley Hills, CA: Sage.

Arimoto, A. (1997). Market and higher education in Japan. *Higher Education Policy*, *10*(3), 199-210.

Ballantyne, C. (1998). Improving university teaching: Responding to feedback from students. In N. Zepke, L.Leach, & A.Viskovice (Eds.), *Adult learning cultures, challenges and choices in times of change* (pp.155-165). Wellington, New Zealand: WP Press.

Boise, R. (1992). *The new faculty member*. San Francisco: Jossey Bass.

Braskamp, L., Brandenburg, D., & Ory, J. (1984). *Evaluating teaching effectiveness: A practical guide.* Thousand Oaks, CA: Sage.

Braskamp, L., & Ory, J. (1994). *Assessing faculty effectiveness*. San Francisco: Jossey Bass.

Burden, P. (2005). Student ratings: Meaningful voices or meaningless ritual? *On Cue*, *13*(2), 16-26.

Canagarajah, S. (2006). TESOL at forty: What are the issues? *TESOL Quarterly*, *40*(1), 9-34.

Cashin, W. (1990). Student ratings of teaching: Some recommendations for use. *IDEA paper No. 22. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.* 1-5.

Cashin, W. (1995). Student ratings of teaching: The data revisited. *IDEA paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.* 1-9.

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

Chen, Y., & Hoshower, L. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education*, *28*(1), 71-89.

Dickey, D., & Pearson, C. (2005). Recency effect in college student course evaluations. *Practical Assessment, Research and Evaluation*, *10*(6). Retrieved December 3, 2007, from <http://pareonline.net/getvn.asp?v=10&n=6>

Dunegan, K., & Hrivnak, M. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*(3), 280-303.

Goodman, R. (2005). W(h)ither the Japanese university? An introduction to the 2004 higher education reforms in Japan. In J. Eades, R. Goodman & Y. Hada (Eds.), *The "Big Bang" in Japanese higher education* (pp. 1-31). Melbourne: Trans Pacific Press.

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, *5*(4), 419-434.

Kitamura, K. (1997). Policy issues in Japanese higher education. *Higher Education*, 34, 141-150.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin (Ed.), *Changing practices in evaluating teachers* (pp. 45-70). Bolton, MA: Anker Publishing.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

Ministry of Education, Culture, Sports, Science and Technology. (2004). *FY2003 white paper on education, culture, sports, science and technology*. Retrieved November 17, 2005, from <www.mext.go.jp/english/news/2004/05/04052401.htm>

Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education*, *10*(1), 57-73.

Norris, J (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *The Modern Language Journal*, 90, 576-583.

Patton, M. (1996). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Schmelkin, S., Spencer, K., & Gellman, E. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, *38*(5), 575-593.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, 39, 40-43.

Simpson, P., & Siguaw, J. (2000). Student evaluation of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22*(3), 199-213.

Spencer, K., & Schmelkin, L. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education*, *27*(5), 397-409.

Stiggins, R., & Duke, D. (1988). *The case for commitment to teacher growth: Research on teacher evaluation*. Albany, NY: State University of New York Press.

Taut, S., & Brauns, D. (2003). Resistance to evaluation: A psychological perspective. *Evaluation*, *9*(3), 247-264.

Theall, M., & Franklin, J. (1990). Student ratings in the context of complex evaluation systems. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (pp. 17-34). San Francisco: Jossey Bass.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction. In M. Theall, P. Abrami, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (pp. 45-56). San Francisco: Jossey Bass.

Tsurata, Y. (2003). Globalisation and Japanese higher education. In R. Goodman & D. Phillips (Eds.), *Can the Japanese change their education system?* (pp. 119-151). Oxford: Symposium Books.

Wachtel, H. (1998). Student evaluation of college teaching effectiveness: A brief overview. *Assessment and Evaluation in Higher Education*, *23*(2), 191-213.

Yamada, R. (2001). University reform in the post-massification era in Japan: Analysis of government education policy for the 21st century. *Higher Education Policy*, 14, 277-291.

Yamamoto, S. (2005). Government and the national universities: Ministerial bureaucrats and dependent universities. In J. Eades, R. Goodman, & Y. Hada (Eds.), *The "Big Bang" in Japanese higher education* (pp. 94-105). Melbourne: Trans Pacific Press.