UNIVERSITY STUDENTS EVALUATING ENGLISH LANGUAGE TEACHERS IN JAPAN: IS OBJECTIVITY LOST IN TRANSLATION?

Gregory Minehane
(minehane@meijo-u.ac.jp)
Meijo University, Japan

Abstract

During several years of a liberal arts English program, carrying out student evaluations of teaching (SETs) was mandatory. Scores obtained from the SETs gradually came to be used, controversially, for summative purposes. A retroactive examination of the SET data from 224 1st and 2nd semester classes encompassing 9742 student surveys showed that student satisfaction correlated highest with rapport. The current study examines contemporary issues in SET research including social and psychological influences, generalizability, and bias. With a focus on Japan and the type of factors that may influence students rating teaching of communicative English classes at university, the case is made for caution regarding the procuring of SETs for summative purposes.

1 Introduction

At face value, incorporating student evaluations of teaching (SETs) into a language program with the aim of better meeting the needs and wants of students is a commendable undertaking. Employing some form of survey, commonly including a question asking students to rate their overall level of satisfaction with a class, is a frequent occurrence. Perhaps because the very use of SETs in universities nowadays is so widespread, program developers may add it to their checklist of 'must-do's' automatically, without due consideration of the controversy which surrounds this extensively researched area of teaching. Lacking an understanding of the differences between formative and summative evaluation, procuring SETs can inadvertently do damage to the very program that directors are trying to improve. Fundamental to the debate is also the question of whether SETs are an appropriate tool at all for gauging the efficacy of an individual teacher's instruction in effecting measurable learner outcomes (Spooren et al., 2013).

In the field of applied linguistics, numerous researchers have sought to define what constitutes good teaching behaviors and instruction - for the most part they have had a great deal of trouble coming to a consensus. Medgyes (2001) found that there are just too many variables that need to be taken into account when trying to paint a picture of the ideal instructor. In spite of, or perhaps because of this conundrum, we have placed the onus on students to tell which of us outshines his or her colleagues. Under what conditions can this process have validity? Moody states that "if student evaluations are to serve as a measure of teaching proficiency, they must prove to reflect the actual amount of learning that occurs during the course" (1976, p.454). In other words, a highly-rated teacher should be a catalyst of greater learning. If research could show this empirically, credence could be given to the argument that SETs are a valid means of assessing a teacher's performance. Without a measurable greater improvement in learning, how much meaning should we attach to the comparatively higher SET score of one instructor or method of instruction over another?

Because of a common lack of understanding of the distinction that should be made between implementing formative and summative evaluation processes, poorly formulated and administered SETs can be counterproductive (Popham, 1988). SET's with formative goals can be done to elicit student feedback to then help tweak programs or individual instructors – they are used to help the teacher teach more effectively. Summative SETs are done at the end of a course or program to judge the efficacy of the course or its instructor – these are linked to monetary rewards. What seems to be the case, however, is that educational environments inevitably suffer from the same pressure that most businesses do. That is, high customer or student satisfaction is demanded by stakeholders to show that the system-in-use has the approval of the end-user. SETs may begin with formative aims but as pressure on the administration increases (say from a budgetary or even pedagogical point of view) – then a commensurate need grows for objective data to show that student satisfaction is also high or rising. The corollary, of course, is that teachers with higher satisfaction ratings on SETs are given more classes (hence overall average satisfaction increases). Systems of employment may even be created that reward higher SET scoring teachers with better paying jobs. SETs are no longer formative in nature (although they may try to maintain this veneer) and are now summative, less for the benefit of students or instructors, more for the benefit of program administrators. SETs are emotive, not only because instructors feel that the typical outcome bears much resemblance to a popularity contest but also because they feel that the results don't always reflect the effort they have put into their teaching (Burden, 2007). As the results may also be tied to promotion and reward within an institution, we may be implicitly teaching our teachers the importance of becoming more popular.

1.1 Actual Versus Perceived Learning

Clayson carried out a meta-analysis of 42 studies of SET that were published between 1953 and 2007, controlling the variable of 'learning' by only incorporating data in the analysis when actual results on validated tests were available. He found that none of the studies that met these criteria and that took place after 1990, showed a positive significant relationship between measured learning and the SET score. He did however, find that "the student's satisfaction with, or perception of, learning is related to the evaluations they give" (2009, p.26). His research suggested that as the statistical analysis of data became more stringent in newer studies, and the more objective the definition of learning became, correlation between SET and student improvement increasingly ceased to exist, and even at times became negative.

Students may in fact be unable to distinguish between actual and perceived learning. If SETs are only measuring perceived learning then sociological and psychological influences are likely to be more persuasive on student survey responses than the effects of real learning. Clayson (2009) acknowledges that much of the research that has been done in SET involves examining students who study in a broad range of disciplines. For the most part, these disciplines do not have the validated diagnostic tools available to English language teachers. For example, in some studies, individual instructor's grades may be used to measure learning. SET research may also include students evaluating their own learning through questions such as "Do you feel you learned a lot in this course?"

1.2 Generalizability and Bias in SET Research

Spooren et al. attempts to summarize through meta-validation much of the peer-reviewed research published since 2000. Examining a database of 160 papers, they conclude that SET research has very little generalizability because research is carried out in unique settings with mostly unique, non-standardized instruments. They lament the contradictory research results on SET, stating that this must be "(at least partly) due to the great variety of methods, measures, controlling variables, SET instruments, and populations used in these studies" (2013, p.629).

A great variety of factors appear to influence how students evaluate and perceive their instructors. These include physical attractiveness (Hamermesh and Parker, 2005), gender (Kogan et al., 2010), class size (Bedard and Kuhn, 2008), teacher's race (McPherson et al., 2009) and even sexual orientation of the instructor (Ewing et al., 2003). Over the period of

interaction with a teacher, Shevlin (2000) asserts that students form a global opinion or evaluation of their instructor which approximates how they perceive the teacher's charisma. This general appraisal then guides them in making the ratings of individual questions in a survey and has a great influence on them; greater than their perception of whether the teacher has taught them well or not.

1.3 Human Psychology

To increase our understanding of what may be influencing students in their response to SET we need to lastly examine research in the field of psychology, particularly research in social influence. With the exception of the SET phenomenon it is usually the teacher who evaluates the student, not vice visa. The norm of reciprocation is a pervasive social force and, according to Kelln and Ellard (1999) pushes us to achieve equity and trust in our relationships. If over the semester students have been made to feel that they are performing well (they perceive their English is improving), it seems likely that they would reciprocate with a positive SET for their teacher. Beran and Violato (2005) do not discuss their results in terms of behavioral psychology, but do find a significant relationship between students' higher expectations of success in the course with higher SETs.

Another well-documented human trait is that of liking. Both teachers and students will try to create and nurture meaningful social relationships. While the teacher-student relationship has its own complexities, it is still subject to the laws of social exchange. Abrami et al. (1997) argues that personal positive views of instructors lead to more positive SET scores. When teachers ask students to complete SETs for them, they seem more likely to receive a favorable response if they have personalized their relationship and are liked by the students. Implicit theories of personality based on the classic 1950's work of Gestalt psychologist Solomon Asch also support this view (see Asch, 1951).

2 Methodology

2.1 Background and Procedures

General English classes from six faculties who were part of the liberal arts program of a large private university in central Japan were managed collectively. Middle-management of the English language program was interested, among other factors, in the level of satisfaction students had with their English classes. The survey was created for non-research purposes. Student satisfaction scores were collated and reported back to instructors along with student comments and suggestions for improving teaching. Administering such student surveys was

compulsory for all 60 full-time and part-time instructors in the liberal arts program (comprising of 32 native and 28 non-native English instructors). SETs in the liberal arts program were completed by students after taking the unified end-of-semester exam in both first and second semesters. Up to 20 minutes was allocated for this although it was usually completed in less than 10 minutes.

2.2 Participants

The liberal arts English program itself ran from 2005 to 2014 and was subsequently disbanded in favor of a return to a policy of general English being the province of individual faculties. SETs were not done in every year of the program and due to privacy restrictions, data, including raw survey data, was not always accessible. Data analyzed and reported in this paper is from the first and second semesters of the 2009/2010 academic year. All students in first and second years completed the SETs. 224 individual English classes containing 5033 students in first semester and 224 classes with 4709 students from the second semester were included in the analysis.

2.3 Materials

The survey used with the liberal arts program contained 16 questions. Students were questioned in Japanese and for the most part all responses were made in Japanese. Questions 1-11 were answered with 5-point rating scales. It should be noted that while the verbal descriptors of the scales varied from question to question, they all presented ‘5’ as being the most predisposed to a particular feeling (strongly so) and ‘1’ being the least disposed. It was also the case that scores for teachers were regularly reported as the summation/average of these scores, so their treatment as being interchangeable here has precedence. Only Q.1 should be thought of as failing to meet the requirement for statistical linearity but for ease of explanation has been included in the correlation matrix reported below. For questions 11 through 16, spaces were provided for students to write their own answers. Note that these open-response questions and student comments are not the focus of the current research and have not been included for discussion or analysis.

3 Results and Analysis

Analysis of students’ responses to the SETs using multivariate analysis of variance measured the correlation between responses to different questions on the survey. All analyses were carried out using JMP 8 statistical software. Of particular interest to the current research was

students' response to Question 11 which read, 'Please rate your overall satisfaction level regarding this class.' Correlation estimates by Pairwise methods were examined to see how responses to this question correlated with students' answers to other questions on the survey. The type of statistical analysis performed is analogous to the kind of analysis needed for delineating factors – items in a survey for research purposes that could be found to be measuring the same construct. Supplementary to the exploratory factor analysis described above, Cronbach's alpha was calculated to check item internal consistency. For the entire set of questions (excluding Q. 1), Cronbach's alpha was .8849, n= 9742, telling us that the items were very closely related as a whole (see http://www.ats.ucla.edu/stat/spss). Five separate analyses were performed: 1). All data, first and second semesters, native and non-native instructors, 2). First semester only, native and non-native taught classes, 3). Second semester only, native and non-native taught classes, 4). First and second semesters, native instructors only, and 5). First and second semesters, non-native instructors only.

Table 1. All data (n=9742) (aggregate of 1st & 2nd semesters, native & non-native instructors), correlation matrix of SET questions (Q.1-11) with emphasis on correlations found with overall student satisfaction (Q.11)

	Q.1	Q.2	Q.3	Q. 4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10	Q.11
Q.1 How much time do you spend doing English activities outside the classroom? (i.e. homework, reading etc.)	1.000	.0890	.1292	.1042	.2026	.1416	.1616	.1501	.2392	.2664	**.1638**
Q.2 Does the lesson start at the correct time and finish at the correct time?		1.000	.3753	.3428	.2392	.2865	.3603	.3108	.2529	.2546	**.3436**
Q.3 Is the teacher's voice loud enough, and is the speech clear and easy to understand?			1.000	.5168	.3250	.3366	.5209	.5078	.3424	.3814	**.4908**
Q.4 Has the teacher presented all manner of materials used in class in a clear and understandable fashion?				1.000	.3721	.3684	.4911	.4607	.3570	.3922	**.4677**
Q.5. Did you find the textbooks helpful for you to learn English?					1.000	.3809	.4306	.4156	.6076	.5415	**.4961**
Q.6 Does the teacher make good use of the textbooks, as well as other supplementary materials in class?						1.000	.4222	.3611	.3336	.3150	**.4040**
Q.7 Does the teacher make an effort to be understood and to maintain the students' interest?							1.000	.6566	.4819	.5130	**.6060**
Q.8 Has the teacher established a good relationship with the students?								1.000	.5171	.5745	**.6396**
Q.9 Do you feel that your English has improved any thanks to having taken this class?									1.000	.7037	**.5758**
Q.10 Has taking this class helped you to like English better?										1.000	**.6231**
Q.11 Please rate your overall satisfaction level regarding this class.											**1.000**

Analyses found that Q. 8 was most highly correlated with Q. 11. Question 8 was 'Has the teacher established a good relationship with the students?' – leading to the finding that the best indicator of high satisfaction with an English class was this element of rapport. Also of note, correlation has increased somewhat in second semester, suggesting even greater alignment between how a student has liked a class (their 'satisfaction with it') and the quality of the relationships that the teacher has been able to establish over the longer period of time. It made no difference whether classes were taught by a native or non-native instructor – satisfaction correlated most highly with the level of relationship created between the instructor and their language class.

Table 2. Correlations of SET questions (Q.1-10) with overall student satisfaction (Q.11) for 1st & 2nd semesters and native & non-native instructors

	1st Sem. (n=5033)	2nd Sem. (n=4709)	Native Instructor (n=4818)	Non-native Instructor (n=4924)
Q.1 How much time do you spend doing English activities outside the classroom? (i.e. homework, reading etc.)	.1768	.1573	.1619	.1600
Q.2 Does the lesson start at the correct time and finish at the correct time?	.3274	.3591	.3206	.3551
Q.3 Is the teacher's voice loud enough, and is the speech clear and easy to understand?	.4274	.5093	.4977	.4835
Q.4 Has the teacher presented all manner of materials used in class in a clear and understandable fashion?	.4476	.4896	.4601	.4648
Q.5. Did you find the textbooks helpful for you to learn English?	.4922	.4994	.4855	.4996
Q.6 Does the teacher make good use of the textbooks, as well as other supplementary materials in class?	.3718	.4368	.3962	.3941
Q.7 Does the teacher make an effort to be understood and to maintain the students' interest?	.5871	.6253	.6043	.5999
Q.8 Has the teacher established a good relationship with the students?	.6129	.6662	.6342	.6449
Q.9 Do you feel that your English has improved any thanks to having taken this class?	.5812	.5704	.5780	.5707
Q.10 Has taking this class helped you to like English better?	.6095	.6371	.6234	.6218
Q.11 Please rate your overall satisfaction level regarding this class.	1.000	1.000	1.000	1.000

While outside the scope of the present study, it warrants reporting that the very highest of intercorrelations among question items in the analysis of 'All data' was found between Q. 9 and Q. 10 (.70) (see Table 1). It is a commonly held belief that a strong connection exists

between feelings towards a subject area and perceived learning so this is also not a surprising result.

4 Discussion and Conclusion

Analysis of the current data has shown us that there is a high degree of intercorrelation among all survey item responses. Students appear to rate teachers consistently, somewhat regardless of the specifics of the questions they were asked. In other words, a student's response to any one question on the survey gives us a clue as to how they would answer the other questions. The data suggest, for example, that a student wouldn't be likely to express great satisfaction with a class, yet be critical of the volume of a teacher's voice. It would seem that there is a more of an all-or-nothing type of evaluation in progress; if one score is high, they will all be high, and if one is low, they will all be low. This result fits best with Shevlin's (2000) notion of students appraising teachers based on a global-like perception of the teacher's charisma. Student ratings simply may not reflect the actual effectiveness of the teaching.

Results of the current study show that students' class satisfaction ratings correlate highest with perceptions of rapport – how good the instructor's relationship with the class is thought to be. The findings point to the role of rapport as an important indicator of student satisfaction; if a student thinks the instructor has made friends with students, it is likely the class satisfaction rating will also be high. The rapport indicator has recorded higher correlations with satisfaction than even a student's perception of how much their English ability improved due to the class (Q. 9). This result was found consistently across each of the statistical analyses that were performed. Rapport was the best indicator of satisfaction for 1st semester, 2nd semester, native instructors, non-native instructors and for the entire data set.

In an English communication class in Japan there may be as few as 15 to 25 students. Teachers endeavor to engage students in a variety of topics both situational and personal. Perhaps it isn't so unusual for teachers to have remembered their student's names shortly after the semester has started. From a psychological and social perspective, the phenomenon of liking and reciprocity will logically have more impact the greater personal contact becomes. The small numbers of students in a class, and the nature of communication and conversation classes, represent the ideal conditions indicated by Abrami et al. (1997) for SET scores to be positively affected by the relationships that teachers have created over the course of the year. While further research may be necessary, factors such as these may be influencing student responses on SET greater than presently accounted for.

The university teaching environment in Japan may be further complicated when the maturational level of the SET respondents is taken into account. In Australia, for example, mature age students (those over the age of 30 years old) account for more than a quarter of the university student population (www.justlanded.com/english/Australia/Australia-Guide/Education/Introduction).
Many studies in SET, do not take into account student's ages, although I think we can expect student opinions to mature quite significantly as they get older (see Spooren, (2010) who found older students gave higher SETs). Almost without exception, in Japan, students in first year are 18 or 19 years old. The developmental and consequent maturational level of Japanese students should not be overlooked as a possible source of influence on their disposition and ability to complete SETs. While empirical research regarding this assertion is lacking, it seems safe to say that a culture of *kouhai* rating *sempai* isn't a strong point of Japanese culture. Students enter university with little experience in rating teaching, psychometrics and because of their young age, even life experience itself.

Issues detailed by Clayson (2009) are not easily set aside: students' perception of their own learning may be measurable but their true progress is much harder to quantify and then to subsequently attribute to the efforts of a single instructor. Methodically speaking, there are many issues in measuring student learning and there are psychological, social, cultural and developmental influences on the student sample which consequently affect how SET data and studies in SET can be interpreted. At the very least these issues should make us question the generalizability of the results of SET studies, in particular those that do not relate the students' evaluation of teaching back to an equivocal standard regarding student learning.

By recognizing that SETs may not actually be measuring what they purport, and taking into account the numerous factors discussed above concerning the environment in which SETs are administered in the communicative English teaching Japanese classroom, a lack of objectivity regarding instructors and instruction in general, is not surprising. When research in support of SETs has been carried out in educational settings vastly different from the conditions that most English teachers face in Japan, there appear to be sufficient issues to cast doubt on the legitimacy of relying on SET results to make managerial or policy decisions concerning teacher employment. On SET surveys conducted and analyzed in the current research, the greatest indicator of student satisfaction was the rapport the teacher was able to establish with his or her students. Current findings add to the corpus of research which suggests that SETs have dubious value in the summative evaluation of teachers.

References

Abrami, A. P., d'Apollonia, S. & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R.P. Perry & J.C. Smart (Eds.), *Effective teaching in higher education: research and practice.* (pp. 321-367). New York: Agathon Press.

Asch, S. (1951). Effects of group pressure on the modification and distortion of judgement. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp.177-190). Pittsburg CA: Carnegie.

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*, 253-265.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education*, *30*, 593-601.

Burden, P. (2007). An analysis of teachers' negative reactions to the implementation of student evaluation of teaching surveys in Japan. In C. Gitsaki (Ed.), *Language and Languages: global and local tensions* (pp 307-327). Newcastle UK: Cambridge Scholars Publishing

Clayson, D.E. (2009) Student evaluations of teaching: Are they related to what students learn? A Meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, *1*, 16-30.

Ewing, V. L., Stukas, A. A., & Sheehan, E. P. (2003). Student prejudice against male and lesbian lecturers. *The Journal of Social Psychology*, *143*, 569-579.

Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369-376.

Kelln, B.R.C. & Ellard, J.H. (1999). An equity theory analysis of the impact of forgiveness and retribution on transgressor compliance. *Personality and Social Psychology Bulletin*, *25*, 864-872.

Kogan, L., Schoenfeld-Tacher, R., & Helleyer, P. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*, 623-636.

McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*, 37-51.

Medgyes, P. (2001). When the teacher is a non-native speaker. In M. Celce-Murcia (Eds.), *Teaching English as a second or foreign language*, (pp. 415-427). Boston: Heinle & Heinle.

Moody, R. (1976). Student achievement and student evaluations of teaching in Spanish. *The Modern Language Journal, 60,* 454-463.

Popham, W.J. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education, 1*, 269-273.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*, 397-405.

Spooren, (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36,* 121-131.

Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluations of teaching: The state of the art. *Review of Educational Research, Vol. 83, 4,* 598-642.