Shiken: JALT Testing & Evaluation SIG Newsletter, 1 (1) April 1997 (p. 12 - 19)

Student evaluation of teachers: Professional practice or punitive policy?

T. L. Simmons

Teachers are evaluated by various methods. At this point in time the manner in which teachers are evaluated in Japan is undergoing change in some ways due to social and legal pressures (Shiozawa, 1995) Labour Standards Laws in this country, for example, actually contain the directives to refrain from dismissal unless there is reasonable need to do so (Sugeno, 1992, pp. 395-412). Although these laws are not necessarily being applied strictly in the work place and in the courts, we are seeing a greater sensitivity to human rights that will undermine many overtly discriminatory practices we see commonly now. In the place of invidious discrimination enacted with impunity we will see greater sophistication in the manner in which decisions are made to obstruct professional activities by some and to dismiss others for responding to the ethical considerations of their professional commitments. A fairly common sort of evaluation that may actually be used for the best intentions but often facilitates the most common abuses is the use of student opinion in the decisions that effect teachers.

In spite of the evidence

In 1993, H. T. Tagomori, at the University of San Francisco, did his Ed. Doctorate on instruments used for student evaluation of faculty. He established that the assessment used by universities and colleges to appraise a professor's teaching effectiveness were conducted by evaluation through instruments they design, borrow, or adapt from other universities and colleges. The reliability of the instruments used is generally unknown. A comprehensive content analysis of faculty evaluation instruments has not been conducted. As a result, faculty members in higher education may be evaluated with flawed evaluation instruments, conceivably leading to unfair assessment of their teaching performance.

". . . faculty members in higher education may be evaluated with flawed evaluation instruments, conceivably leading to unfair assessment of their teaching performance."

Tagomori did a content analysis of 4,028 evaluation items contained in the 200 evaluation instruments analyzed. His analysis revealed 54.6% of the items were ambiguous, unclear and/or subjective. Another 24.5% of the items did not correlate with classroom teaching performance. Altogether a total of 79.1% of the items were either flawed or did not identify with teaching performance. The content analysis also revealed 58% of the 200 evaluation instruments contained responses to evaluation items that were ambiguous, positively skewed or negatively skewed. Based on a frequency-count recording and frequency distribution of the data, the conclusion for this study

is that evaluation instruments used in their present form are unreliable. It would seem clear that current SETEs (Student Evaluations of Teacher(s) (Effectiveness)) for evaluating teaching performance at universities and colleges must be systematically revised. And I would add that the purpose for their use must also be constructively addressed.

What exactly is being evaluated?

O'Connell and Dickinson (1993) stated that while it is well known that factors other than the instructors' teaching influences student ratings of instruction, not all of the sources have yet been identified. In addition, the correspondence between the amount of student learning and student ratings has not been clearly established. O'Connell and Dickinson also noted that although SETEs are one of the most common processes used in evaluating promotion, tenure and other benefits, their reliability is so low that they should not be used for judging individual performance (Stedman, 1983).

Who is being evaluated and who is evaluating?

Student evaluations of teacher effectiveness (SETEs) are, at best, nothing more than evaluations of the students' perceptions of the teachers' effectiveness – at best. It should be intuitively apparent to most that opinions expressed are subject to a great many variables that may have little or nothing to do with evaluating the teachers' ability to teach. The problems of SETEs are multitude and do not categorically represent a professional endeavour within education. SETEs must be justified in design, administration, analysis, interpretation and application of the interpretation. There is no published research to show that this is a task institutions in Japan are willing or prepared to do.

Variables that effect students' opinions

Class type

Class type is usually defined by such characteristics as lecture, combined lecture discussion, and laboratory. Wigington, Tollefson and Rodriguez in a study involving 5,483 evaluations of 242 different classes (1989) showed that instructors with discussion classes had higher overall ratings than other class types and smaller classes showed a higher rating than larger within a class type. There were no differences among class types in comparisons of class rank (lower, upper or graduate). But they did show that in some types the more experienced instructors got the lower ratings and the least experienced got the lowest in others – instructor rank is apparently interactive with class type. The Wigington, Tollefson and Rodriguez study on the interaction of class type and instructor gender (1989) showed that women got better ratings in lecture, discussions and laboratory

classes but received lower scores in lecture/discussion classes.

Instructor rank & reputation

Instructor rank can be differentiated as graduate teaching assistant (GTA) and various levels of professor. The GTAs position of part-time teacher in what are usually lower division classes may be filled by adjunct lecturers in Japan although the adjunct faculty usually have graduate degrees or doctorates. Research by Wigington, Tollefson and Rodriguez (1989) into the interaction of class size and instructor rank showed that some ranks were greatest at the small class size, decreased in mid class size and somewhat greater in large class size – a 'U' shaped profile. In other cases, the class size showed a inverse relationship with the instructors rank – bigger classes, lower ratings. There were some interesting variations in Wigington, Tollefson and Rodriguez's study of the interaction of instructor rank and instructor gender (1989). At both ends of the ranking system, graduate assistants and full-professors, women got higher than men of the same rank, but the middle ranks of assistant professor and associate professor showed that there were no difference for men and women. In their enquiries into the interaction of level and instructor rank, Wigington, Tollefson and Rodriguez (1989) found significant variation; it was somewhat directly proportional for assistant professors and professors but 'U' shaped for graduate assistants (higher at both ends of the graph than in the middle) and 'bell' shaped for associate professors. Comparatively, GTAs had higher scores in lower level classes than did professors and lower scores for upper division classes. Full-professors did not evince the highest scores at any level.

The reputation of the instructor among the students may also be a factor. Wigington, Tollefson and Rodriguez (1989) found that students often take a class on the basis of the instructors' reputation. When this happened, the ratings were higher for lecture and laboratory class types. The interaction for reputation and rank showed the highest variation for professors. They received the highest scores from students who took classes for the instructors' reputation but they were also the lowest among the ranks when the students did not take the class because of the instructor's reputation. GTAs showed the least variation (Wigington, Tollefson & Rodriguez, 1989).

Students' conceptions of performance

In judging a performance, the raters (students in this case) may be biased by their belief as to what constitutes a job well done. If that is not enough of a quandary, psychometrically oriented studies have largely ignored the fact that performance judgements are made with incomplete information – the students must recall or infer performance behaviour (Kishor, 1995). We are asked to believe that the student has a realistic idea of how teaching a foreign language is done well and they are doing it from memory. They are, in short, being asked questions that they have not

necessarily considered and asked to answer them accurately after the fact in a matter of perhaps one hour.

Gender

Many major reviews of student evaluations conclude that gender does not have a significant effect (Seldin, 1993; Marsh & Dunkin, 1992). However Basow (1995) provides sufficient data from the literature to show that many of these studies examines only the main effects. Gender varies with the gender of the student as well as the teacher, the gender typing of disciplines, status of the professor (e. g. tenured vs. non-tenured), teaching styles, student year, student grade point average, student grade expectations, number of years teacher has taught, the hour the class is taught, student perceptions of the teachers speech, thought stimulation, non-repetition, and overall rating. The point that Basow makes is that for individual teachers, the results of the interaction of numerous variables, negligible alone, may be significant if the influences occur simultaneously. "Anyone using student evaluations should have a sophisticated understanding of how gender variables may operate in such ratings." (Basow, 1995).

Class size

The effect of class is uncertain (Wigington, Tollefson & Rodriguez, 1989). Feldman (1978, 1984) found no significant effect and Smith and Glass (1980) and Whitten and Umble (1980) found that instructors of smaller classes have higher ratings.

Using a demarcation of small as 25 or less, mid sized as 26-49, and large as any class over 50, Wigington, Tollefson and Rodriguez, (1989) found that the interaction of instructor gender and class size produces better ratings for women in small classes and men in large classes. The interaction of class level (lower, upper or graduate division) and size showed higher ratings for instructors in upper division, large classes in comparison with mid-sized classes.

Class level

Classes can be ranked by divisions, lower, upper, and graduate. The research on the effects of class level on ratings is ambiguous (Wigington, Tollefson & Rodriguez, 1989). Romeo and Weber (1985) and others found that there were no difference for ratings of instructors of higher or lower classes. Cranton and Smith (1986) and others found that instructors of upper level classes received higher ratings.

Wigington, Tollefson and Rodriguez (1989) found that while instructors of lower level classes received lower ratings than those of higher level classes, there was no significant difference between men and women in lower levels. However, in upper division or graduate courses men got higher ratings than women.

A self-fulfilling prophecy

A concern to address in the use of student evaluations is the impact the act of evaluation has on the students' perceptions of the teachers and on the teachers themselves.

There are biases in evaluating a person's personality, performance and competence – biases that can lead to flawed information gathering strategies that are self fulfilling (Harris, 1994). A self-fulfilling prophecy as defined by Merton (1948) basically means that an incorrect perception, belief or definition of a set of circumstances can evoke behaviour that makes the incorrect perceptions or beliefs come true.

In the composition of the SETEs the administrators bring their own expectations about the teachers to the procedure. These expectations profoundly effect the way they design the SETEs and the information gathering strategies they use.

In clinical psychology in the study of interpersonal expectancy effects or behavioural confirmation, the problem of making incorrect diagnosis supported by presumptive questioning strategy is a serious ethical issue that remains a central focus. Observers, no matter how well trained and how ethical, will carry out their evaluations based on incorrect hypothesis.

Snyder and Swann (1978), in a classic study, gave subjects a list (personality profile) describing either an extroverted personality or an introverted personality and then asked them to choose 12 questions from a longer list that would best allow them to test the hypothesis for the profile they received for a target person. Analysis demonstrated a heavy emphasis on hypothesis-confirming strategies.

The process of question selection and the process applying those questions to the evaluation of a person's behaviour are difficult for well trained clinicians to perform objectively – the situation of untrained students and administrators and teachers is even more problematic.

When an administration or administrator has decided that teachers fit certain stereotypes or engage in certain types of behaviour – negative or constructive – the administrator will select hypothesis confirming questions for the students to answer.

For example, students are asked if the teacher is humourous, do they like the teacher, does the teacher stimulate or encourage them, is the teacher enthusiastic and dynamic – an entire battery of subjective parameters appear on SETEs that lead the students to believe that the teacher must conform to certain and possibly irrelevant behavioural parameters that actually have a different appeal to each individual student.

As a student answers objective and subjective questions – what will a student rely on – what

they feel confident they can answer or what they are unsure about?

The nature of objective questions present certain problems. How can a student know whether a teacher is well prepared – how do they assess preparedness? How can a student evaluate a teacher's expertise in their field – if they know so much about the field why are they the student? Yet students will give answers to these types of questions which shows that even when they do not have a defensible point of view – they will give an opinion. This is not the way to solicit informed opinions.

Additionally, it is not the students' opinions that have necessarily been solicited; they will be answering someone else's questions without having given the matter any thought until the point in time when they are supposed to 'evaluate' the teacher.

The administrators' perceptions of the teachers can also profoundly effect the teachers' perceptions of their own effectiveness. Teachers who are told that they are teaching poorly because they don't appeal to the parameters the students are asked to rate on the SETEs may in fact be teaching at a competent level but the administrations' input from the tainted SETEs can be amplified by insisting that they are accurate and show the teacher to be less than competent.

And through all of this is the underlying belief that the process of education is predominantly the sole burden of the teacher. The assumption that the teacher is primarily responsible completely colours the students' attitude and the evaluation designer's intent. In this scenario, there is no room for a well rounded evaluation of the students, the management, the facility, the social pressures and inhibitions – a long list of variables is ignored.

"the underlying belief[s] that the process of education is predominantly the sole burden of the teacher. . . . In this scenario, there is no room for a well rounded evaluation of the students, the management, the facility, the social pressures and inhibitions - a long list of variables is ignored."

In real classrooms

Students' subjective opinions can be so varied that the overall results are untrustworthy. Students who are specifically shown that certain SETE parameters have been fulfilled may still evaluate related criteria ambivalently. Students may pointedly refer to a teacher's physical characteristics or manner in very negative or positive terms and judge the teacher on the basis of these characteristics – as if teachers who are not aesthetically acceptable are rendered less capable of teaching.

The entire process of SETEs becomes a convenient matter of picking and choosing what serves to comply with the original hypothesis of the SETE designer/administrator rather than actually engaging in an honest evaluation. This means the evaluation is rather like a shopping list of

potentially conforming characteristics that further the administrators' personal biases.

A proposed paradigm

Adapted from Arnoult and Anderson (1988) to provide for a better paradigm for the evaluation of teacher effectiveness in the academic environment so as to reduce an evaluator's biases: (a) gather as much evidence as possible, (b) employ multiple evaluators who have different view points and interests, (c) vary the observational circumstances to provide for different emphasis in the environment, (d) review video tapes for greater accuracy, (e) compare the criteria on balance sheets to establish evidence for and against an evaluation, (f) solicit an explanation of the results and the subsequent conclusions made by evaluators to reveal gaps in reasoning. This paradigm constitutes constructive advice for the evaluations we make of others in a professional setting.

This type of evaluation is an example of a structured attempt at measuring professional competence with regard for the various facets of the evaluating process which is primarily designed to inform the teachers rather than to judge them – a philosophy that serves better to encourage improvement rather than to punish.

References

Arnoult, L. & Anderson, C. A. (1988). Identifying and reducing causal reasoning biases in clinical practice. In D. C. Turk & P. Salovey (Eds.), Reasoning, inference, and judgment in clinical psychology (pp. 209-232). New York: Free Press.

Basow, S. A. (1995). Student evaluations of college professor: When gender matters. Journal of Educational Psychology. 87, 656-665.

Darley, J. M., Fleming, J. H., Hilton, J. L., & Swann, W. B. (1988). Dispelling negative expectancies: The impact of interactional goals and target practices on the expectancy of the confirmation process. Journal of Experimental Social Psychology, 24, 19-36.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. Research in Higher Education, 9, 199-242.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. Research in Higher Education, 21, 45-116.

Harris, M. J. (1993). Information gathering strategies in social perception. Unpublished manuscript, University of Kentucky, Lexington. Cited in Harris, 1994.

Harris, M. J. (1994). Self-fulfilling prophecies in the clinical context: Review and implications for clinical practice. Applied and Preventive Psychology, 3 (3) 145-158.

Kayne, N. T. & Alloy, L. B. (1988). Clinician and patient as aberrant actuaries: Expectation-based distortions in assessment of covariation. L. Y. Abramson (Ed.) Social cognition and clinical psychology: A synthesis, (pp. 295-365). New York: Guilford Press.

Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgement: consequences for the validity of student ratings of instruction. Research in Higher Education, 36 (2) 177-195.

Marsh, H. W., & Dunkin, M. J. (1992). Student's evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.). Higher education: Handbook of theory and research. (Vol. 8. pp. 143-233). New York: Agathon Press.

Merton R. K. (1948). The self-fulfilling prophecy. Antioch Review, 8, 193-210.

Nielsen, R. S. (1993). The impact of the 1985 reform legislation on the formative evaluation practices of one central Illinois school district. Doctoral Dissertation, University of Illinois at Urbana-Champaign (in Harris, 1994:148).

O'Connell, D. Q., & Dickinson, D. J. (1993). Student ratings of instruction as a function of testing conditions and perceptions of amount learned. Journal of Research and Development in Education, 27 (1) 18-23.

Sackett, P. R. 1982. The interviewer as hypothesis tester. The effects of impressions of an applicant on interviewer questioning strategy. Personnel Psychology, 35, 789-804.

Seldin, P. (1993, July 21). The use and abuse of student ratings of professors. The Chronicle of Higher Education, p. A40.

Shiozawa T. (1995). The change of the Monbusho guidelines and their impact on language education. Paper. JALT 95, Nagoya Japan. Reprinted in PALE Newsletter,(1996) 2, 1.

Smith, M. L. & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. American Education Research Journal, 17, 419-433.

Snyder, M., & Campbell, B. (1980). Testing hypothesis about other people: the role of the hypothesis. Personality and Social Psychology Bulletin, 6, 421-426.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. Journal of Personality and Social Psychology, 36, 1202-1212.

Snyder, M., & Thomasen, C. J. (1988). Interactions between therapists and clients: Hypothesis testing and behavioural confirmation. In C. D. Turk & P. Salovey (Eds.), Reasoning, inference and judgement in clinical psychology. New York: The Free Press.

Stedman, C. H. (1983). The reliability of teaching effectiveness rating scale for assessing faculty performance. Tennessee Education, 12 (3) 25-32.

Sugeno K. (1992). Japanese Labour Law, (Leo Kanowitz, Translator) Tokyo: University of Tokyo Press.

Swann, W. B., Jr., & Ely, R. J. (1984). A battle of wills: Self-verification versus behavioural confirmation. Journal of Personality and Social Psychology, 46, 1287-1302.

Swann, W. B., Jr., & Giuliano, T. 1987. Confirmatory search strategies in social interaction: How, when, why, and with what consequences. Journal of Social and Clinical Psychology, 5, 511-524.

Tagomori, H. T. (1993). A content analysis of instruments used for student evaluation of faculty in schools of education at universities and colleges accredited by the national council for accreditation of teacher education. Unpublished Ed. Doctorate dissertation. University of San Francisco.

Turk, C. D., & Salovey, P. (Eds.) 1988., Reasoning, inference and judgement in clinical psychology. New York: The Free Press.

Wigington, H., Tollefson, N. & Rodriguez, E. (1989). Student's ratings of instructors revisited: Interactions among class and instructor variables. Research in Higher Education, 30 (3) 331-344.

Whitten, B. J., & Umble, M. M. (1980). The relationship of class size, class level and core vs. non-core classification for class to student ratings of faculty: Implications for validity. Educational and Psychological Measurement, 40, 419-423.