

# Profiling Teacher/Teaching Using Descriptors Derived from Qualitative Feedback: Formative and Summative Applications

Daphne Pan · Gary S. H. Tan · Kiruthika Ragupathi ·  
Krishna Booluck · Rita Roop · Yuen K. Ip

Received: 21 August 2007 / Published online: 23 September 2008  
© Springer Science+Business Media, LLC 2008

**Abstract** Considerable work has been done on student evaluation of teaching/teachers, but reservations remain about its use for summative purposes. Student ratings are not universally accepted as being reliable, nor can they provide really meaningful information. Qualitative comments can provide a better understanding but they tend not to be user-friendly from lack of structure and connectedness. This study attempts to devise a method for ‘quantifying’ students’ comments to increase their usefulness in complementing/confirming ratings. The quantified results enable the profile construction of what students regard as an effective/ineffective teacher, and enable identification of strengths and weaknesses. Our findings counter some commonly held assumptions, including those which held that high ratings are dependent on small class size and ‘dumbing down’ of courses and the consequent expectation of high grades. The findings also indicate that students value teaching quality more than teacher characteristics, suggesting their ability to make valid judgments about teaching effectiveness.

**Keywords** Profiling teacher/teaching · Student ratings · Qualitative feedback · Students’ written comments · Text analysis · Formative evaluation · Summative assessment

---

D. Pan (✉) · G. S. H. Tan · K. Ragupathi · K. Booluck · R. Roop · Y. K. Ip  
Centre for Development of Teaching and Learning, National University of Singapore,  
10 Kent Ridge Crescent, Singapore, Singapore 119260  
e-mail: cdthead@nus.edu.sg

D. Pan  
Department of English Language and Literature, National University of Singapore, Singapore,  
Singapore

G. S. H. Tan  
Department of Computer Science, National University of Singapore, Singapore, Singapore

Y. K. Ip  
Department of Biological Sciences, National University of Singapore, Singapore, Singapore

## Introduction

### Student Feedback

Student feedback refers to students' evaluation of a teacher's characteristics and teaching. In higher education, student feedback is indispensable in teacher evaluation. Well constructed teaching evaluation instruments render valid and reliable information on teaching effectiveness (Gallagher 2000). The evaluation can be quantitative, providing ratings on a graded scale and/or comprise a qualitative section which solicits comments. The rationale for this open-ended approach is to encourage frank responses from students and allow them to highlight other areas of the teacher's strengths and weaknesses that the student ratings might have missed.

### Student Ratings

To date, numerous papers have been written on the study of students' evaluation of teaching (Marsh and Dunkin 1997). Although research began in the thirties and forties (Remmers 1949; Guthrie 1954), student ratings in those times were usually obtained on a voluntary basis. However, they have now become a requirement at most institutions, and their use has shifted from the formative—which encourages instructional development—to the summative where they play an important part in personnel decisions. While there has been extensive research on the reliability and validity of student ratings in the past decades (Ory 2000), studies on students' qualitative feedback, especially on its value as a summative tool, is scarce (Lewis 2001).

In the eighties, research investigated the validity of student ratings and what these ratings really meant (Millman 1981); whether they really reflected how good a teacher was, or merely the teacher's personality, entertainment skills, or any of several extraneous factors not associated with teaching (Centra 1987). Many studies in the eighties (Aleamoni and Hexner 1980; McKeachie 1979; Marsh 1984) looked at the validity of student ratings by checking them against how much students learnt from a particular professor. These studies, generally lower-level courses involving many teachers, found that student ratings were reasonably correlated with student learning.

In the nineties, there was a steady accumulation of evidence suggesting the misuse or overuse of ratings data (Franklin and Theall 1989). Debates ensued over the validity of student ratings (for instance in *Chronicle of Higher Education and Change*). However, these did not invalidate the potential of ratings data as useful information about teaching performance (Centra 1981, 1993; Marsh 1987; Theall and Franklin 1990, 1999), and such data continued to be used as a primary source of evidence in teaching evaluation systems in many universities. Administrators often look to student ratings as a convenient, and often the only measure of teaching quality. The ratings of teaching may be used judgmentally to inform decisions about hiring, contract renewal, tenure, promotion and funding of courses (Kember and Wong 2000).

The renewed emphasis on student-centred learning approaches highlighted a need for new evaluation methods which were not biased towards conventional teacher-centred approaches (Abrami et al. 2007). In fact, more researchers and writers have focused on the application of evaluation and ratings research in practice, with the goal of developing "comprehensive evaluation systems" (Arreola 2000). At the same time, new concerns have been raised about another aspect of the ratings and evaluation picture—the rapidly

increasing use of on-line systems (Johnson and Sorenson 2004). From instance, Feldman (2007) argues that the emergence of private or for-profit online ratings has increased faculty's concern about the misuse of ratings data.

### Multi-Dimensionality of Student Ratings of Teaching Performance

Within the literature, there has been abundant research on the dimensionality of student ratings of teaching performance (Marsh 1991; Marsh and Dunkin 1997; Abrami et al. 1990; 1996; Feldman 1997). According to Feldman (1997), teaching consists of many elements and therefore, instruments of teaching evaluation should capture this multidimensionality. Marsh and Dunkin (1997) also argued that a multidimensional approach to student evaluation of teaching would provide more useful feedback to teachers than a single score summary. It would also help student in their selection of courses as well as administrators in personnel decisions.

However, there still remains much debate as to whether a single score or multidimensional profile score is more appropriate. According to Apodaca and Grad (2005), this is complicated by the different formative and the summative goals of teaching evaluation; whilst formative goals require precise and specific feedback as a guide to the improvement of teaching performance; summative goals need an overall parsimony that facilitates the evaluation and judgment making process. Although effective teaching is multidimensional, there are differences across instruments used to rate specific dimensions that underlie effective teaching (Abrami et al. 1996). Therefore, caution should be exercised in using student ratings of specific teaching dimensions indiscriminately for summative purposes.

Most of the studies on the multi-dimensionality of teaching have operated from a quantitative perspective. Factor analyses or similar techniques have been frequently used to determine the different dimensions of teaching, and to establish the relative contributions of each dimension to effective teaching (Marsh 1982, 1983; Ryan and Harrison 1995; Harrison et al. 2004). The Student Evaluation of Educational Quality (SEEQ) questionnaire (Marsh 1982) is among the more well-known quantitative instruments used to analyse teaching profiles. In the development of the SEEQ, faculty were asked to rate the importance and potential usefulness of the items (as a basis for feedback) to be included in the questionnaire (Marsh and Dunkin 1997). Marsh (1987) identified nine factors of instruction, which he argued could be used for summative evaluations. Marsh and Bailey (1993) also used multivariate profile analysis to show that each teacher has a characteristics profile on the nine SEEQ scores (e.g., high on organization and low on enthusiasm). Rosenfield et al. (2005, cited in Abrami et al. 2007), in a study of learning environments, performed factor analyses on a list of 40 items separately for students and instructors. The results showed that students who rated the learning environment as largely student centered, rated their instructors as more effective in helping them learn as opposed to their peers who found the learning environment as more teacher centered. Feldman (1997), in an updated study of his research since the 1970s, identified 28 dimensions of teaching by looking at the average correlation between students and faculty in their judgement of these components. To determine which of these dimension had a greater impact on the overall teaching effectiveness, Feldman (1997) compared the magnitudes between the overall evaluations by students and their ratings of each of the specific characteristics of their teachers.

## Students' Written Comments

Some studies have suggested that open-ended feedback may provide a more 'top of the mind' approach about how students really perceive the teacher. From instance, such feedback may allow a teacher to know what to maintain, improve or change in his teaching, and can help to shape what takes place in the classroom (Wirtz 2004). Braskamp and his colleagues (cited in Marsh 2007) examined the usefulness of students' written comments and their relation to rating items for teacher evaluation.<sup>1</sup> While it was found to be useful for self-improvement and making course changes, Ory and Braskamp (1981) suggested that written comments might be deemed too subjective for important personnel decisions.

Furthermore, Lin et al. (1984) noted that student comments were not easily summarised due to the effort required as well as their idiosyncratic nature, such that it might be more appropriate to simply return the comments to teachers along with summaries of student ratings. Lewis (2001) recommended using conceptual frameworks and matrices to analyse students' written comments, though its application across the campus can be restrictive and time-consuming.

Hence, this study was undertaken to devise an effective method for quantifying qualitative comments to examine its usefulness in complementing student ratings as a summative tool. While we analysed students' evaluation of teaching as a multi-dimensional phenomenon, our approach was different from those of other studies mentioned above as the items used to evaluate teaching effectiveness were derived from the student comments themselves rather than from a pre-selected set of attributes. The circumstances under which student feedback is administered at NUS, and the availability of appropriate commercial text analysing software like *SPSS Text analysis for Surveys* (STAS) render this study possible and timely.

## Student Feedback at NUS

Student feedback was first introduced at NUS in 1992, and was made part of the official summative assessment of teaching quality in 1996. There are several important features concerning student feedback at NUS:

A common instrument is used across all faculties, with an option of two faculty/discipline-specific items, if needed. The instrument is administered online at the end of each semester.

Students evaluate the "module" and the "teacher (and teaching)" separately and they also have the option of nominating the teacher for an award. [Note: NUS operates on a modular system whereby students read "modules" (equivalent to "courses" elsewhere) and gain modular credits (MCs) that reflects the workload required for each module. In general, a module has 50 contact hours, usually worth 4 MCs and lasts one semester.]<sup>2</sup>

The instrument solicits students' rating of the instructor on seven different items based on a 5-point scale (5 being the highest score). An eighth item invites rating of the overall effectiveness of the teacher. The students' written comments on the positive and negative

<sup>1</sup> Refer to Marsh (2007) for a list of Braskamp's relevant articles.

<sup>2</sup> The faculties of Dentistry and Medicine at NUS do not follow the modular system and hence do not collect data at the end of each semester.

aspects of the teacher's characteristics and teaching are obtained through two open-ended questions: "What are the teacher's strengths?" and "What improvements would you suggest to the teacher?" Students are also required to rate the level of difficulty and expected grade of that module on a 5-point scale, and give an overall module score. However, the focus of this paper is on evaluating teaching effectiveness rather than the modules themselves.<sup>3</sup>

Respondent anonymity is guaranteed to encourage frank responses. Faculty members can only access their respective feedback a week after the release of examination results, and the ratings and comments cannot be traced back to particular students. As such, the issue of students giving higher ratings to teachers in order to obtain better grades does not arise. [Note: the schools of Business and Computing have made the student feedback system transparent which allows faculty members to access their peers' feedback.]<sup>4</sup>

To encourage a higher rate of return, students who responded get earlier access to their grades than those who did not. While this raises the possibility of a bias in sampling, with bright and highly motivated students (expecting better grades) being more likely to respond, the overall student feedback response rate was 79.7%, which clearly indicates that the sample was large enough to be representative of the student community.

### Problems Encountered

At NUS, student feedback is one of three components used to assess a faculty member's teaching performance; the others being peer review reports and teaching portfolio. At present, there seems to be an overemphasis on student ratings. In some departments, the ranking of teaching staff is based predominantly on student ratings and, in some cases, the ranking serves as an important criterion in determining the outcome of tenure and promotion exercises and the nomination and/or selection for teaching awards. However, this reliance on the use of student ratings as a summative tool has been accompanied by a certain degree of discomfort among faculty members. Some of the major concerns are:

- factors such as class size; level of difficulty and expected grade can affect student ratings;
- student ratings may be nothing more than a popularity contest with friendly, humorous, easy-grading teachers emerging as winners every time;
- students cannot make valid judgements concerning the teacher and teaching because of their immaturity, lack of experience and capriciousness.

These factors are neither unique to NUS nor novel, and have been previously suggested by Aleamoni (1987). A pertinent and more important concern is the lack of meaning attached to the scores. This may result in individual and subjective interpretation that is misused when making important decisions.

Although a section for qualitative comments is included in the student feedback instrument, examining, analysing and interpreting these comments is one of the most

---

<sup>3</sup> Teachers who taught more than one module were identified to be in the highest or lowest 20% cohort based on the average of their score on overall teaching effectiveness rather than the average module score.

<sup>4</sup> Perry et al. (1974) suggested that prior expectations of teaching performance could influence ratings of professors, and that alternating certain kinds of information caused drastic changes in personality evaluations. However, research by Bejar and Doyle (1976) about this expectation-evaluation relationship showed that students were capable of rating their instructors independently of expectations held prior to the course, and that student interest in the subject matter was often independent of what instructors did.

difficult tasks for faculty members (for self-improvement), and teaching evaluation committees (for assessment of teaching quality). This is because, unlike scaled items (as in student ratings), the written comments have no built-in structure and are usually presented as a series of random, unconnected statements about the teacher and the teaching. It is difficult for the human mind to make sense of such information (Lewis 2001).

Even though faculty members rarely study the written comments received, putting in the extra effort can be beneficial. From instance, a teaching award winner at NUS described how he achieved excellent ratings (compared to the low scores he had when he first started teaching) by making use of students' written comments about his teaching performance (Wirtz 2004). However, there is no organised effort at NUS to analyse and extract meaning from student written feedback in a systematic way. Therefore, there is a need to design a meaningful way to present the information in a comprehensible manner. This constitutes the major thrust of this project. The large data sets—including ratings and written comments—that are captured electronically, and the high rate of response (~80%), offer specially favourable circumstances at NUS for analysing students' written comments in a quantitative fashion using STAS.

## Objectives

This study was undertaken to achieve the following objectives:

1. We aimed to gain insight into the value system of NUS students regarding teaching and learning, by establishing profiles of positive and negative descriptors, detailing what students regarded as an effective (based on the highest 20% cohort) and an ineffective teacher (based on the lowest 20% cohort).
2. Since the use of descriptors could be affected by the command of English and the cultural background of the respondents, efforts were made to interpret the meaning of some important descriptors through analysing their shared responses with other descriptors.
3. An attempt was made to present graphically the profiles based on descriptors of what students regarded as an effective or an ineffective teacher, so that a comparison could be made with the profile of an individual faculty member in order to reveal the strengths and weaknesses of the faculty member in relation to the NUS community.
4. Since two individuals receiving comparable student ratings might have totally different approaches to teaching, we aimed to understand these ratings by analysing the faculty member's positive and negative descriptor profiles, and understand the basis on which individual faculty members were given a certain score.
5. We analysed the written comments of teaching award winners to elucidate why NUS students regarded them as effective teachers with outstanding teaching performance.
6. Finally, we hoped to infer whether interpreting students' written comments quantitatively would be a meaningful way to generate data for formative and summative assessment of teaching.

## Methodology

Qualitative research involves the study of actual situations as they happen naturally, without predetermined constraints on outcomes (Patton 1990). Due to the open-ended nature of the two questions used in the NUS student feedback instrument—"What are the

teacher’s strengths?” and “What improvements would you suggest to the teacher?” the qualitative feedback obtained satisfy the criteria of qualitative research.

The primary methods of collecting data for qualitative research are interviewing, observing and studying printed materials (Wulff and Nyquist 2001). Students’ written comments are printed materials and can also be considered as a form of interviewing (Lewis 2001). Thus, in essence, we have made an attempt in this study to generate quantitative data from qualitative research.

### Data Collection

At the end of the academic year 2005–2006, we obtained student feedback data, both qualitative and quantitative, for the highest 20% cohort ( $N = 278$ ) and the lowest 20% cohort ( $N = 278$ ) of faculty members based on student ratings on overall teaching effectiveness for module levels 1,000–3,000 (undergraduate), 4,000 (honours) and 5,000 (graduate). From faculty members who had taught more than one module, the average score of overall teaching effectiveness of all the modules taught, was used to determine if they were in these two cohorts. Upon establishment of these two cohorts, data on ratings and written comments were obtained (based on the module which had the highest or lowest score for the highest 20% cohort and lowest 20% cohort, respectively) for each teacher. In addition, data on class size, expected grade and level of difficulty were obtained. We also solicited data on teaching award winners between academic years 2003/2004 and 2005/2006 ( $N = 34$ ) for comparison.<sup>5</sup>

### Data Processing

We started out by constructing a list of positive descriptors according to the seminal study by Lowman (1996), but when those descriptors were used to scan students’ written feedback, results obtained were sub-optimal. NUS students apparently did not use many of those descriptors (Lowman 1996) to comment on their teachers’ teaching. Therefore, STAS was used to perform a frequency-based classification to sieve through the recurring keywords and phrases from the written comments. Lists of positive and negative descriptors were constructed based on comments for the questions “What are the teacher’s strengths?” and “What improvements would you suggest to the teacher?” respectively. Subsequently, efforts were made to manually examine 10% of the written feedback from both cohorts to check whether any additional descriptors had to be added. The list was then organised into semantic groupings of keywords to form descriptor categories. From example, “has deep knowledge” and “very knowledgeable” were grouped under

---

<sup>5</sup> NUS has two university-level teaching awards: the Annual Teaching Excellence Award (ATEA) and the even more highly selective Outstanding Educator Award (OEA). The first recognizes excellence in teaching in the year under review as evidenced by high quality teaching practices (including classroom teaching, assessment practices, module development, range of teaching, supervision of student projects) and activities in professional self-development. The OEA recognizes someone who *is not only a consistently excellent teacher but who has made exceptional contributions to the educational culture and practices at NUS*. In addition to teaching practices and professional development, the OEA also considers educational leadership (within NUS), educational impact (national/international), scholarship in teaching, teaching materials adopted by external parties, other relevant activities). Nominations may be made by self or peers and the selection process is a very rigorous one involving both Faculty Teaching Excellence Committees and the University Teaching Excellence Committee, based on information provided by student feedback, peer reviews and self reports through the Teaching Portfolio.

“knowledgeable” while “poor lecture notes” and “incomplete notes” went under “ineffective notes”. The final list of generated categories was used to facilitate analysis.

Since STAS only gave the overall frequency count of each descriptor category, a simple program was developed, using the programming language *Python*, to calculate the descriptor frequency for each individual faculty member. The STAS output (descriptor categories) and class size were supplied as input to *Python* to obtain the frequency count of each descriptor category for each individual. To account for the varying class size (resulting in different number of comments for each teacher) the frequency count was normalised to a class size of 100 students.

## Statistics

Results were presented as mean  $\pm$  standard deviation (S.D.). In many cases, large S.D. Were obtained due to the huge range of data, but because of the large sampling size ( $N$ ), standard errors of mean (S.E.M.) were usually small. Levene’s Test for Equality of Variances was used to examine variances between means. The difference between two means was compared by Students’  $t$ -test (two-tailed). The difference between multiple means was compared by one-way analysis of variance (ANOVA) followed by Bonferroni multiple range test. Differences with  $P < 0.05$  were regarded as statistically significant. Pearson’s correlation was used to analyze the strength, direction (positively or negatively related) and significance of relationships where applicable.

## Results

### Ratings on Overall Teaching Effectiveness, Class Size, Expected Grade and Level of Difficulty for the Highest 20% Cohort and Lowest 20% Cohort

The average ratings on overall teaching effectiveness received by the highest 20% cohort were significantly higher than that received by the lowest 20% cohort (Table 1), while the average class size (students per class) taught by faculty members in the latter was significantly larger than that in the highest 20% cohort. Students taught by the highest 20% cohort had an average expected grade significantly higher than those taught by the lowest 20% cohort (Table 1). Average ratings on level of difficulty of modules taught by the highest 20% cohort and the lowest 20% cohort were comparable.

There were negative correlations between the rating on overall teaching effectiveness and class size for faculty members in both the highest 20% cohort and the lowest 20% cohort (Table 1). By contrast, positive correlations between the rating on overall teaching effectiveness and the expected grade were obtained for both the highest 20% and lowest 20% cohorts. On the other hand, both the highest 20% and the lowest 20% cohorts demonstrated no significant correlations between the rating on overall teaching effectiveness and the level of difficulty of the module (Table 1).

### Ratings on Overall Teaching Effectiveness, Class Size, Expected Grade and Level of Difficulty for Teaching Award Winners

The teaching award winners were compared against faculty members in the highest 20% cohort, as we would expect them to be distinctly different from the lowest 20% cohort.



**Table 1** Average ratings of overall teaching effectiveness, class size (with range in parenthesis), expected grade and level of difficulty, and correlation between some of these items, obtained by the highest 20% cohort ( $N = 278$ ) and lowest 20% cohort ( $N = 278$ ), as defined by the ratings of overall teaching effectiveness, and by the teaching award winners ( $N = 34$ )

	Highest 20%	Lowest 20%	Teaching award winners
Overall teaching effectiveness (A)	$4.42 \pm 0.19^b$	$3.51 \pm 0.32^a$	$4.51 \pm 0.17^b$
Class size (B)	$87 \pm 128^a$ (11–966)	$119 \pm 137^b$ (11–1008)	$177 \pm 206^b$ (11–955)
Expected grade (C)	$4.09 \pm 0.28^a$	$3.96 \pm 0.27^b$	$4.03 \pm 0.19^{ab}$
Level of Difficulty (D)	$3.74 \pm 0.33$	$3.78 \pm 0.35$	$3.67 \pm 0.32$
Correlation between a and B	$-0.183^{**}$	$-0.186^{**}$	$-0.438^{**}$
Correlation between a and C	$0.303^{**}$	$0.147^*$	$0.074$
Correlation between a and D	$-0.052$	$-0.084$	$-0.237$

Results are presented as mean  $\pm$  SD

Means for (A), (B), or (C) not sharing the same letter are significantly different (ANOVA;  $P < 0.05$ )

\* Significantly correlated,  $P < 0.05$  (2-tailed)

\*\* Significantly correlated,  $P < 0.01$  (2-tailed)

From teaching award winners, the average rating on overall teaching effectiveness was greater than that for the highest 20% cohort. Incidentally, the average class size taught by the teaching award winners was significantly greater than that taught by faculty members in the highest 20% cohort. As for the expected grade and level of difficulty, the average scores were comparable between the teaching award winners and the highest 20% cohort (Table 1).

Surprisingly, unlike the highest 20% and lowest 20% cohorts, teaching award winners did not reflect significant correlation between ratings for overall teaching effectiveness and expected grade. However, as with those two cohorts, there was a significant negative correlation between ratings for overall teaching effectiveness and class size, and there was no significant correlation between the former and the level of difficulty of the module (Table 1).

### Descriptor Profiles of an Effective Teacher and an Ineffective Teacher in NUS

A profile of what NUS students regarded as an effective teacher was established based on the 20 most frequently used positive descriptors obtained from the highest 20% cohort of teachers, as defined by their high ratings on overall teaching effectiveness (Table 2). Out of these 20 positive descriptors, only 7 were on teacher characteristics, and out of those 7, “approachability” was ranked first. With respect to teaching performance, “interesting” appeared with the highest frequency. The teacher’s ability to “explain” (4th), “deliver concepts” (11th), “stimulate thinking” (13th) and encourage “real-life applications” (20th) are also regarded by students as important parameters in evaluating teaching. Surprisingly, only one teacher characteristic (approachability) was positioned among the top 5 descriptors. Thus, it would appear that teaching performance is ranked as being more important than teachers’ personal traits (e.g. Friendly, patient or humorous).

An analysis based on the twenty most frequently used positive descriptors obtained from the lowest 20% cohort, revealed a similar profile. Teachers in this cohort impressed students for being “knowledgeable” (1st), while being “interesting” fell to the 5th position

**Table 2** The top twenty most frequent positive descriptors and negative descriptors obtained from the highest 20% cohort ( $N = 278$ ) or the lowest 20% cohort ( $N = 278$ ), as defined by the student rating on the overall teaching effectiveness

Rank	Highest 20%		Lowest 20%	
	Positive descriptors	Negative descriptors	Positive descriptors	Negative descriptors
1	Interesting	Ineffective lecturing	Knowledgeable	Ineffective lecturing
2	Approachable	Pace of teaching	Approachable	Unclear
3	Clarity	Ineffective notes	Ability to explain	Poor elocution
4	Ability to explain	Ineffective use of examples	Clarity	Ineffective notes
5	Effective teaching	Time management	Interesting	Pace of teaching
6	Knowledgeable	Ineffective slides	Patient	Time management
7	Willing to help	Poor questioning	Friendly	Ineffective use of examples
8	Aids understanding	Poor elocution	Aids understanding	Not interesting
9	Friendly	Poor explanation	Willing to help	Ineffective slides
10	Patient	Ineffective use of concepts	Effective teaching	Poor explanation
11	Delivery of concepts	Not interesting	Delivery of concepts	Difficulty in understanding
12	Humorous	Unclear	Good lecture notes	Ineffective use of concepts
13	Stimulates thinking	Difficulty in understanding	Effective questioning	Problems with tutorials
14	Effective use of examples	Problems with tutorials	Stimulate thinking	Poor questioning
15	Encouraging	Not enough real-life applications	Effective use of examples	Unhelpful
16	Effective questioning	Problems with assessments	Humorous	Not detailed enough
17	Engaging	Unhelpful	Encouraging	Not enough real-life applications
18	Good lecture notes	Not detailed enough	Informative	Disorganised
19	Concise	Problem with readings	Detailed	Unprepared
20	Real-life applications	Simplify explanations/concepts/terms	Real-life application	Problems with assessments

in this profile. Interestingly, “good lecture notes” was ranked more highly in this profile (12th) than in the profile obtained from the highest 20% cohort (18th) (Table 2).

“Ineffective lecturing” was the most frequently used negative descriptor obtained from both the highest 20% and lowest 20% cohorts, indicating that lecturing skills were regarded highly by NUS students. The profiles of an ineffective teacher obtained from these two cohorts were similar, despite different rankings of various negative descriptors. From instance, “unclear” was ranked second in the lowest 20% cohort, but it was in the 12th position in the highest 20% cohort profile. In both cases, only one teacher characteristic, i.e. “unhelpful” was listed, while the other 19 descriptors referred to problems associated with teaching performance. It is important to note that “problems with assessments” (20th)

appeared in the profile of an ineffective teacher obtained from the lowest 20% cohort, but was absent from the highest 20% cohort profile (Table 2).

### Comparison of Frequency of Positive and Negative Descriptors Between the Highest 20% and Lowest 20% Cohorts

Indeed, the frequency of all 20 positive descriptors obtained from the highest 20% cohort was significantly greater than those obtained from the lowest 20% cohort (Table 3). Interestingly, the only descriptor which shows no significant difference between these two cohorts is “good lecture notes”.

Likewise, the frequency of all 20 negative descriptors obtained from the lowest 20% cohort was significantly greater than those from the highest 20% cohort (Table 4). The frequency for the descriptor “problems with assessment” was comparable between these two cohorts, but substantially different in ranking (ranked 20th in the lowest 20% cohort and 16th in the highest 20% cohort).

**Table 3** A comparison on the frequencies (frequency per 100 students per teacher; mean  $\pm$  SD) of positive descriptors, based on the ranking of descriptors obtained from the highest 20% cohort, between the highest 20% cohort ( $N = 278$ ) and the lowest 20% cohort ( $N = 278$ ) as defined by the student rating on the overall teaching effectiveness

Rank	Positive descriptors	Frequency	
		Highest 20%	Lowest 20%
1	Interesting	5.65 $\pm$ 5.9	2.02 $\pm$ 3.16*
2	Approachable	5.63 $\pm$ 6.36	3.27 $\pm$ 4.19*
3	Clarity	4.91 $\pm$ 4.69	2.12 $\pm$ 2.5*
4	Ability to explain	4.84 $\pm$ 4.98	2.52 $\pm$ 3.23*
5	Effective teaching	3.88 $\pm$ 4.02	1.41 $\pm$ 1.99*
6	Knowledgeable	3.74 $\pm$ 4.47	3.52 $\pm$ 3.95
7	Willing to help	3.6 $\pm$ 4.34	1.45 $\pm$ 2.05*
8	Aids understanding	3.39 $\pm$ 3.57	1.65 $\pm$ 2.15*
9	Friendly	3.06 $\pm$ 3.73	1.66 $\pm$ 2.67*
10	Patient	2.85 $\pm$ 4.07	2.01 $\pm$ 3.1*
11	Delivery of concepts	2.46 $\pm$ 2.81	1.25 $\pm$ 2.17*
12	Humorous	2.08 $\pm$ 3.39	0.65 $\pm$ 1.61*
13	Stimulates thinking	1.89 $\pm$ 3.13	0.83 $\pm$ 1.53*
14	Effective use of examples	1.43 $\pm$ 2.56	0.75 $\pm$ 1.62*
15	Encouraging	1.31 $\pm$ 2.27	0.59 $\pm$ 1.61*
16	Effective questioning	1.29 $\pm$ 2.43	0.85 $\pm$ 1.69*
17	Engaging	1.16 $\pm$ 2.2	0.24 $\pm$ 0.83*
18	Good lecture notes	1.06 $\pm$ 2.22	0.89 $\pm$ 1.7
19	Concise	1.05 $\pm$ 1.94	0.41 $\pm$ 0.9*
20	Real-life applications	1.03 $\pm$ 2.01	0.48 $\pm$ 1.61*

Values represent means  $\pm$  SD

\* Significantly different from the corresponding highest 20% value ( $P < 0.05$ )

**Table 4** A comparison on the frequencies (frequency per 100 students per teacher; mean  $\pm$  SD) of negative descriptors, based on the ranking of descriptors obtained from the lowest 20% cohort, between the highest 20% cohort ( $N = 278$ ) and lowest 20% cohort ( $N = 278$ ) as defined by the student rating on the overall teaching effectiveness

Rank	Negative descriptors	Frequency	
		Lowest 20%	Highest 20%
1	Ineffective lecturing	4.18 $\pm$ 4.18	0.72 $\pm$ 1.78*
2	Unclear	2.98 $\pm$ 3.62	0.18 $\pm$ 0.89*
3	Poor elocution	2.94 $\pm$ 4.03	0.25 $\pm$ 1.01*
4	Ineffective notes	2.86 $\pm$ 4.6	0.58 $\pm$ 1.89*
5	Pace of teaching	2.26 $\pm$ 3.37	0.62 $\pm$ 1.92
6	Time management	2.04 $\pm$ 2.77	0.37 $\pm$ 1.18*
7	Ineffective use of examples	1.29 $\pm$ 2.23	0.52 $\pm$ 2.48
8	Not interesting	1.24 $\pm$ 1.89	0.21 $\pm$ 0.94*
9	Ineffective slides	1.23 $\pm$ 2.28	0.27 $\pm$ 1.04*
10	Poor explanation	1.21 $\pm$ 1.95	0.24 $\pm$ 1.11*
11	Difficulty in understanding	1.19 $\pm$ 1.85	0.18 $\pm$ 0.68*
12	Ineffective use of concepts	0.96 $\pm$ 1.71	0.23 $\pm$ 0.82*
13	Problems with tutorials	0.84 $\pm$ 2.07	0.17 $\pm$ 0.78*
14	Poor questioning	0.82 $\pm$ 1.74	0.26 $\pm$ 0.97
15	Unhelpful	0.73 $\pm$ 1.58	0.12 $\pm$ 0.43*
16	Not detailed enough	0.67 $\pm$ 1.31	0.11 $\pm$ 0.57*
17	Not enough real-life applications	0.53 $\pm$ 1.65	0.15 $\pm$ 0.65*
18	Disorganised	0.49 $\pm$ 1.06	0.04 $\pm$ 0.29*
19	Unprepared	0.48 $\pm$ 1.53	0.02 $\pm$ 0.2*
20	Problems with assessments	0.46 $\pm$ 1.37	0.13 $\pm$ 0.6

Values represent means  $\pm$  SD

\* Significantly different from the corresponding lowest 20% value ( $P < 0.05$ )

Based on statistical differences between profiles obtained from the highest 20% and lowest 20% cohorts (Tables 3 and 4), it is logical to consider the profile of positive descriptors obtained from the highest 20% cohort as the profile of what students regarded as an effective teacher. Similarly, the profile of negative descriptors obtained from the lowest 20% cohort can be regarded as the profile of what students regarded as an ineffective teacher. With the establishment of these two general profiles, we were able to compare and evaluate the descriptors obtained from individual teacher and to use them subsequently as reference points.

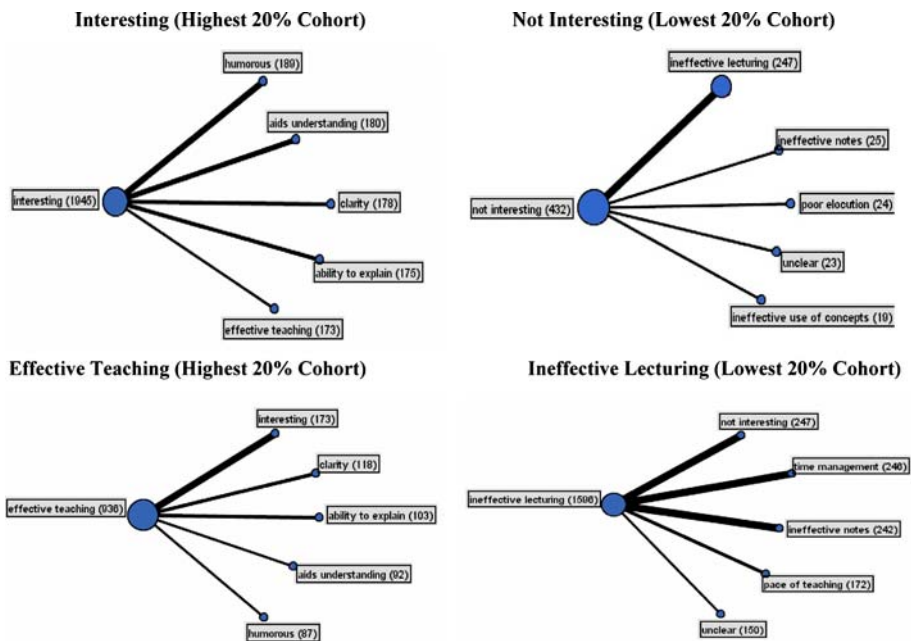
#### Establishing the Meaning of Some Important Positive Descriptors and Negative Descriptors Through Analyses of Shared Responses

Grouping descriptors based on their frequencies can reveal the intricate relationships among them, and shared responses between descriptors can yield information on the meaning of a certain descriptor to students. This is especially important in the analysis of written comments from Asian students because the usage of descriptors could be

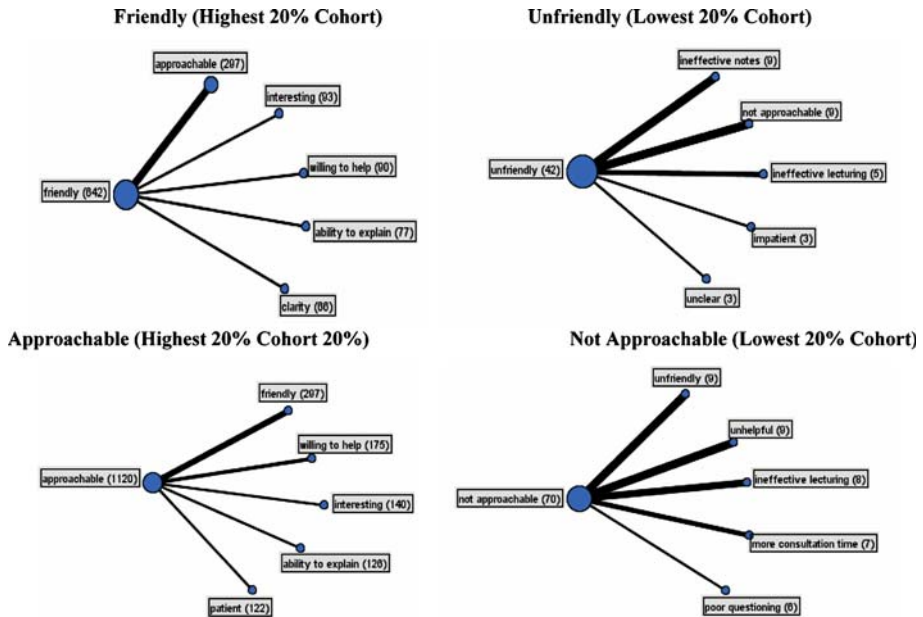
influenced by the respondents' command of English and their diverse cultural backgrounds. This was achieved by using the category web option in the STAS to examine how descriptors overlap. Results are presented as nodes. Each node represents a descriptor, with the size of the node based on the number of responses indicated in brackets for that descriptor. The width of the connecting lines signifies the extent of co-occurrence of connecting nodes.

It should be noted that NUS students used “interesting” not only in frequent association with the teacher characteristic “humorous”, but also teaching qualities like “aids understanding”, “clarity” and “ability to explain” to describe faculty members in the highest 20% cohort (Fig. 1). Similarly, they used descriptors associated with teaching quality in frequent association with “not interesting” to describe faculty members in the lowest 20% cohort (Fig. 1). Of the five most frequently used descriptors associated with “effective teaching” for the highest 20% cohort, four were on teaching qualities, with the exception of “humorous” which had the lowest frequency among the top five highly associated descriptors. As for “ineffective lecturing”, the most frequently used descriptors associated with it were “not interesting”, “time management”, “ineffective notes”, “pace of teaching” and “unclear” (Fig. 1).

More importantly, for teacher characteristics like “friendly/unfriendly”, and “approachable/not approachable”, some of the most frequently associated descriptors were related to teaching qualities, e.g. “ability to explain”, “aids understanding”, “ineffective lecturing” and “poor questioning” (Fig. 2). From this it may be inferred that NUS students



**Fig. 1** The top five descriptors that have the highest shared responses with the positive descriptors “interesting” or “effective teaching” obtained from student written comments of the highest 20% ( $N = 278$ ), and the negative descriptors “not interesting” or “ineffective teaching” obtained from student written comments of lowest 20% cohorts ( $N = 278$ )



**Fig. 2** The top five descriptors that have the highest shared responses with the positive descriptors “friendly” or “approachable” obtained from student written comments of the highest 20% ( $N = 278$ ), and the negative descriptors “unfriendly” or “not approachable” obtained from student written comments of lowest 20% cohorts ( $N = 278$ )

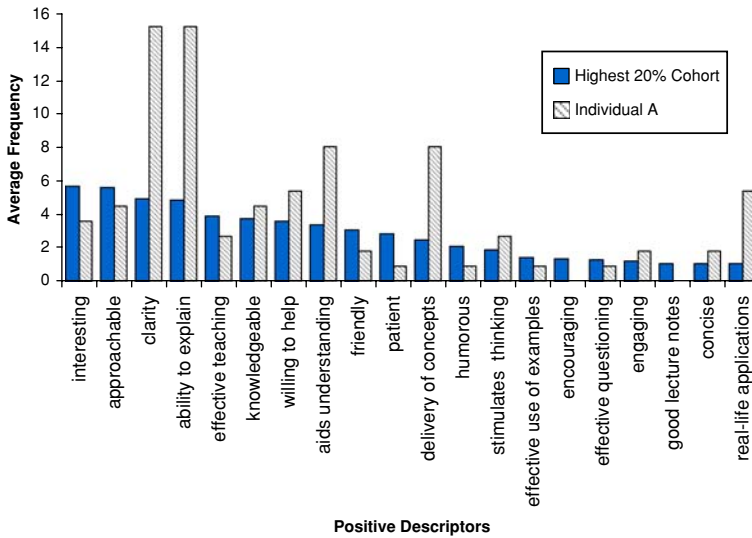
did not over-emphasize teacher characteristics in evaluating the performance of their teachers and teaching.

#### Evaluating an Individual with Reference to the General Profile of an Effective Teacher

Individual A, who had a rating of 4.557 for overall teaching effectiveness, was randomly chosen from the highest 20% cohort. The frequency of the various descriptors listed in the general profile of an effective teacher was obtained from individual A’s student comments and presented together with those obtained from the same cohort graphically (Fig. 3). It is immediately obvious that the strengths of individual a are “clarity”, “ability to explain”, “aids understanding”, “delivery of concepts” and encourage “real-life applications”. The frequency of these 5 descriptors was twice that of the highest 20% cohort. Noteworthy too is that there is no mention of “good lecture notes” in the written comments for individual a and the frequency of descriptors on teacher characteristics (e.g. Friendly, patient, humorous and encouraging) was generally lower than those in the profile of an effective teacher from the highest 20% cohort.

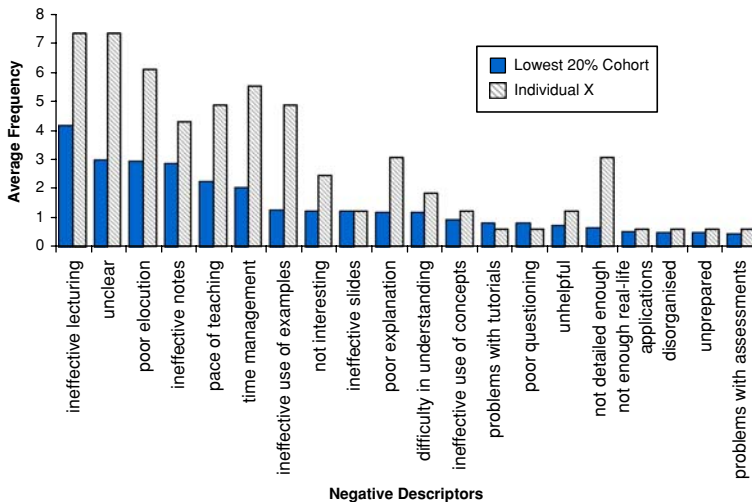
#### Evaluating an Individual with Reference to the General Profile of an Ineffective Teacher

Similarly, individual X who had a rating of 3.384 for overall teaching effectiveness was randomly chosen from the lowest 20% cohort. The frequency of descriptors was obtained



**Fig. 3** A direct comparison between the frequencies (frequency per 100 students per teacher) of the twenty most frequent positive descriptors obtained from the highest 20% cohort (the general profile of an effective teacher;  $N = 278$ ) and the corresponding frequencies (frequency per 100 students) of the same positive descriptors obtained by individual A, who received a student rating of 4.557 for overall teaching effectiveness

according to, and compared graphically with those in the general profile of an ineffective teacher from the same cohort (Fig. 4). Individual X apparently had many problems in teaching, because the frequency of descriptors like “ineffective lecturing”, “unclear”, “poor elocution”, “pace of teaching”, “time management”, “ineffective use of examples”,



**Fig. 4** A direct comparison between the frequencies (frequency per 100 students per teacher) of the twenty most frequent negative descriptors obtained from the lowest 20% cohort (the general profile of an ineffective teacher;  $N = 278$ ) and the corresponding frequencies (frequency per 100 students) of the same negative descriptors obtained by individual X, who received a student rating of 3.384 for overall teaching effectiveness

“poor explanation” and “not detailed enough” were at least twice those in the general profile of an ineffective teacher at NUS.

### Constructing the Personal Profiles or “Fingerprint” of Positive and Negative Descriptors for an Individual

To enable a faculty member to truly know his/her strengths in teaching, and the weaknesses that need to be addressed, it would be useful to construct the personal teaching profile or “fingerprint” of the individual teacher. This was established by interpreting the twenty most frequently used positive and negative descriptors obtained by the individual, which could be quite different from those in the upper profile of an effective teacher and the lower profile of an ineffective teacher.

For individual A1, who was selected randomly from the highest 20% cohort and obtained a score of 4.275 for overall teaching effectiveness, the “fingerprint” revealed that his/her strengths were in being “interesting”, “ability to explain” and “humorous” (Table 5). Descriptors like “provide solutions” and “capture attention” also showed up in the fingerprint profile.

**Table 5** The personal profile and frequencies (frequency per 100 students) of the 20 most frequent positive descriptors and the twenty most frequent negative descriptors of individual “A1”, who had a rating of 4.275 for overall teaching effectiveness, from the highest 20% cohort

Rank	Positive descriptors	Frequency	Negative descriptors	Frequency
<i>Individual A1</i>				
1	Interesting	11.84	Pace of teaching	6.57
2	Ability to explain	10.52	Ineffective notes	2.63
3	Humorous	6.57	Poor questioning	2.63
4	Clarity	5.26	Ineffective use of examples	2.63
5	Friendly	3.94	Time management	2.63
6	Good lecture notes	3.94	More open to suggestions	1.31
7	Detailed	3.94	Difficulty in understanding	1.31
8	Good speaker	2.63	Ineffective lecturing	1.31
9	Aids understanding	2.63	Not enough breaks	1.31
10	Approachable	2.63	Too abstract	1.31
11	Provide solutions	1.31	Reading from notes/slides/books	1.31
12	Stimulates thinking	1.31	Poor elocution	1.31
13	Fun	1.31	Teaching method	1.31
14	Informative	1.31	Ineffective learning	1.31
15	Effective teaching	1.31	Ineffective slides	1.31
16	Patient	1.31	Impatient	1.31
17	Knowledgeable	1.31	Confusing	1.31
18	Capture attention	1.31	Poor explanation	1.31
19	Effective learning	1.31	Hard to grasp	1.31
20	Experienced	1.31	Condescending	1.31



**Table 6** The personal profile and frequencies (frequency per 100 students) of the 20 most frequent positive descriptors of individual “A2”, who had a rating of 4.158 for overall teaching effectiveness, from the highest 20% cohort

Rank	Positive descriptors	Frequency
<i>Individual A2</i>		
1	Approachable	33.72
2	Willing to help	23.25
3	Friendly	16.86
4	Patient	13.95
5	Aids understanding	9.3
6	Effective questioning	6.39
7	Ability to explain	5.23
8	Clarity	4.65
9	Available for consultation	4.06
10	Kind	4.06
11	Effective use of examples	2.9
12	Tries his best	2.32
13	Available	2.32
14	Nice	2.32
15	Extra mile	1.74
16	Enthusiastic	1.74
17	Effective teaching	1.74
18	Concerned	1.74
19	Delivery of concepts	1.16
20	Polite	1.16

In contrast, individual A2, who obtained a similar score of 4.158, showed a different set of strengths: descriptors such as “approachable”, “willing to help”, “friendly”, and “patient” occurred with exceedingly high frequency (Table 6). Probably because of his/her particular approach to teaching, descriptors like “nice”, “concern” and “polite” were picked up in the profile.

An effective teacher would naturally also have weaknesses, and it is apparent from the negative descriptors of individual A1 that “pace of teaching” posed greatest problem to students (Table 5). “interesting” and “pace of teaching”, when taken together imply that he/she was covering materials at high speed because of an eagerness to share interesting information with his/her students.

For individual X1, who was selected randomly from the lowest 20% cohort and obtained a score of 3.464 for overall teaching effectiveness, the “fingerprint” reveals that his/her weaknesses were “ineffective notes”, “ineffective lecturing” and “unclear”, and his/her strengths were “ability to explain, aids understanding” and “interesting” (Table 7). Apparently, students responses were divided over “clarity/unclear”, and “good lecture notes/ineffective notes”, but it would appear that he/she obtained a relatively low score on overall teaching effectiveness because there were more negative opinions on these items.

The usefulness of the fingerprint profiling is readily apparent when one examines the profile of negative descriptors for individual X2, who obtained a rating of 2.730 (Table 8). The most frequently used negative descriptors were “not punctual”, suggesting that he/she

**Table 7** The personal profile frequencies (frequency per 100 students) of the twenty most frequent positive descriptors and the twenty most frequent negative descriptors of individual “X1”, who had a rating of 3.464 for overall teaching effectiveness, from the lowest 20% cohort

Rank	Positive descriptors	Frequency	Negative descriptors	Frequency
<i>Individual X1</i>				
1	Ability to explain	3.7	Ineffective notes	4.07
2	Aids understanding	2.96	Ineffective lecturing	3.7
3	Interesting	2.96	Unclear	3.7
4	Clarity	2.22	Poor elocutionary skills	3.33
5	Good lecture notes	2.22	Ineffective slides	2.96
6	Effective teaching	1.48	Not interesting	1.48
7	Knowledgeable	1.48	Difficulty in understanding	1.48
8	Willing to help	1.48	Pace of teaching	1.48
9	Humorous	1.11	Repetition	1.48
10	Delivery of concepts	0.74	Gives too much/not enough information	1.11
11	Friendly	0.74	More concise	1.11
12	Lively	0.37	Not enough breaks	0.74
13	Tries his best	0.37	Lengthy	0.74
14	Informative	0.37	Confusing	0.74
15	Straight to the point	0.37	Poor explanation	0.74
16	Approachable	0.37	Impatient	0.37
17	Inspirational	0.37	Redundant	0.37
18	Prepared	0.37	Unfriendly	0.37
19	Good questioning	0.37	Incoherent	0.37
20	Punctual	0.37	Dull	0.37

was probably often late for class. In addition, many other descriptors not listed in the profile of an ineffective teacher were captured, e.g. “problem with assessment”, “ineffective lecturing”, “repetition”, “unfair”, “irresponsible”, and “prone to mistakes/errors”.

#### Profiles of Positive and Negative Descriptors of Teaching Award Winners

Written comments on teaching award winners were examined to understand why NUS students regarded them as outstanding teachers. As a group, these award winners had a higher average score on overall effectiveness (Table 1) and also high average frequency of positive descriptors (Table 9). Once again, it is evident that the intellectual dimension was more highly valued by students, with descriptors like “interesting”, “clarity”, “ability to explain” ranking higher than personal attributes like “friendly” or “humorous”. In general, the frequency of the 20 most frequent positive descriptors were higher than those of the highest 20% cohort.

Specifically, the frequency for “interesting”, “aids understanding”, and “engaging” obtained by teaching award winners were significantly greater than those of the highest 20% cohort, indicating that their strengths in these three parameters were especially strong. As for negative descriptors, the frequency for teaching award winners were generally

**Table 8** The personal profile and frequencies (frequency per 100 students) of the 20 most frequent negative descriptors of individual “X2”, who had a rating of 2.730 for overall teaching effectiveness, from the lowest 20% cohort

Rank	Negative descriptors	Frequency
<i>Individual X2</i>		
1	Not punctual	16.79
2	Problems with assessments	12.97
3	Ineffective lecturing	9.16
4	Time management	4.19
5	Problems with tutorials	3.43
6	Poor questioning	3.05
7	Ineffective notes	2.67
8	Repetition	1.52
9	Unfair	1.52
10	Poor explanation	1.52
11	Irresponsible	1.14
12	Unprepared	1.14
13	Unenthusiastic	1.14
14	Not detailed enough	1.14
15	Pace of teaching	0.76
16	Prone to mistakes/errors	0.76
17	Lack of passion	0.76
18	Unhelpful	0.76
19	Knowledge	0.76
20	Ineffective use of concepts	0.76

comparable to those for faculty members from the highest 20% cohort, except for “pace of teaching”, “ineffective lecturing”, “time management” and “poor elocution”, which showed significant differences. Thus, even teaching award winners had weakness in teaching as evidenced by the fact that some students had problems with their fast pace of teaching and poor time management which manifested as ineffective lecturing. However, by comparing the magnitude of frequencies for the top five positive descriptors with those for the top five negative descriptors, it becomes obvious that their strengths in teaching was weighed more heavily than their weaknesses by the students (Table 10).

#### Evaluating Profiles of Individuals Who have Comparable Scores

Two individuals, B and C, who had student ratings of 4.557 and 4.392, respectively, for overall teaching effectiveness, were chosen randomly from the highest 20% cohort. It is immediately obvious from their positive descriptor profiles that they had been ‘rewarded’ for different sets of teaching attributes despite having similar scores on overall teaching effectiveness. The five most frequently used positive descriptors for individual B were: (i) ability to explain, (ii) clarity, (iii) delivery of concepts, (iv) aids understanding, (v) knowledgeable (Fig. 5). More importantly, the frequency (15 per 100 students) obtained by individual B for both “ability to explain” and “clarity” were three times greater than that in the profile of an effective teacher in the highest 20% cohort. This individual is obviously outstanding in teaching performance.

**Table 9** A comparison on the frequencies (frequency per 100 students per teacher; mean  $\pm$  SD) of positive descriptors, based on the ranking of positive descriptors obtained from the teaching award winners, between the teaching award winners ( $N = 34$ ) and the highest 20% cohorts ( $N = 278$ ) as defined by the student rating on the overall teaching effectiveness

Rank	Positive descriptors	Frequency	
		Teaching award winners	Highest 20%
1	Interesting	8.79 $\pm$ 6.97	5.65 $\pm$ 5.9*
2	Approachable	7.01 $\pm$ 6.94	5.63 $\pm$ 6.36
3	Clarity	6.47 $\pm$ 5.19	4.91 $\pm$ 4.69
4	Ability to explain	6.06 $\pm$ 4.56	4.84 $\pm$ 4.98
5	Aids understanding	5.35 $\pm$ 3.48	3.39 $\pm$ 3.57*
6	Effective teaching	5 $\pm$ 3.07	3.88 $\pm$ 4.02
7	Willing to help	3.97 $\pm$ 2.96	3.6 $\pm$ 4.34
8	Friendly	3.12 $\pm$ 3.19	3.06 $\pm$ 3.73
9	Patient	3.06 $\pm$ 4.19	2.85 $\pm$ 4.07
10	Humorous	2.69 $\pm$ 3.46	2.08 $\pm$ 3.39
11	Delivery of concepts	2.69 $\pm$ 2.44	2.46 $\pm$ 2.81
12	Knowledgeable	2.55 $\pm$ 2.26	3.74 $\pm$ 4.47
13	Encouraging	2.35 $\pm$ 3.17	1.31 $\pm$ 2.27
14	Engaging	2.29 $\pm$ 2.63	1.16 $\pm$ 2.2*
15	Effective questioning	1.9 $\pm$ 3.1	1.29 $\pm$ 2.43
16	Stimulates thinking	1.82 $\pm$ 3.16	1.89 $\pm$ 3.13
17	Good lecture notes	1.66 $\pm$ 2.87	1.06 $\pm$ 2.22
18	concise	1.65 $\pm$ 1.74	1.05 $\pm$ 1.94
19	Effective use of examples	1.63 $\pm$ 2.1	1.43 $\pm$ 2.56
20	Real-life applications	1.36 $\pm$ 1.97	1.03 $\pm$ 2.01

Values represent means  $\pm$  SD

\* Significantly different from the corresponding value for the teaching award winners ( $P < 0.05$ )

As for individual C, the five most frequently used positive descriptors were: (i) interesting, (ii) clarity, (iii) ability to explain, (iv) approachable, and (v) engaging (Fig. 6). Individual C had an average frequency of 14.5 for “interesting”, which was 2.5 times greater than the value of 5.65 in the upper profile of an effective teacher in NUS and about 4 times greater than that obtained by individual B. On the other hand, the frequency of “ability to explain” obtained by individual C (5 per 100 students) was only one-third of that obtained by individual B (15 per 100 students). These results indicate that individual C’s ability to get students interested in the subject matter and his/her interpersonal skills contributed to the higher rating on overall teaching effectiveness.

Similarly, two individuals, Y and Z with comparable student ratings for overall teaching effectiveness (3.506 and 3.636, respectively) were randomly chosen from the lowest 20% cohort. Their negative descriptor profiles were analysed in order to gain insight into why they obtained these scores. The five most frequent negative descriptors for individual Y are “ineffective notes”, “problems with tutorials”, “poor elocution”, “time management” and “ineffective lecturing” (Fig. 6).

For individual Z, the five most frequent negative descriptors were: “not approachable”, “too critical of students”, “not interesting”, “poor questioning” and “sarcastic” (Fig. 6).

**Table 10** A comparison on the frequencies (frequency per 100 students per teacher; mean  $\pm$  SD) of negative descriptors, based on the ranking of negative descriptors obtained from the teaching award winners, between the teaching award winners ( $N = 34$ ) and the highest 20% cohorts ( $N = 278$ ) as defined by the student rating on the overall teaching effectiveness

Rank	Negative descriptors	Frequency	
		Teaching award winners	Highest 20%
1	Pace of teaching	2.31 $\pm$ 2.64	0.62 $\pm$ 1.92*
2	Ineffective lecturing	1.62 $\pm$ 1.9	0.72 $\pm$ 1.78*
3	Ineffective notes	1.11 $\pm$ 1.86	0.58 $\pm$ 1.89
4	Time management	0.99 $\pm$ 1.46	0.37 $\pm$ 1.18*
5	Poor elocution	0.99 $\pm$ 1.57	0.25 $\pm$ 1.01*
6	Ineffective use of examples	0.58 $\pm$ 0.85	0.52 $\pm$ 2.48
7	Difficulty in understanding	0.52 $\pm$ 1.05	0.18 $\pm$ 0.68
8	Poor questioning	0.48 $\pm$ 0.8	0.26 $\pm$ 0.97
9	Ineffective slides	0.48 $\pm$ 1.16	0.27 $\pm$ 1.04
10	Problems with tutorials	0.32 $\pm$ 0.84	0.17 $\pm$ 0.78
11	Problems with projects	0.31 $\pm$ 0.96	0.09 $\pm$ 0.47
12	Unclear	0.31 $\pm$ 0.69	0.18 $\pm$ 0.89
13	Unhelpful	0.3 $\pm$ 0.66	0.12 $\pm$ 0.43
14	Ineffective use of concepts	0.29 $\pm$ 0.55	0.23 $\pm$ 0.82
15	Poor explanation	0.29 $\pm$ 0.73	0.24 $\pm$ 1.11
16	Not interesting	0.28 $\pm$ 0.47	0.21 $\pm$ 0.94
17	Not enough real-life applications	0.26 $\pm$ 0.49	0.15 $\pm$ 0.65
18	Confusing	0.24 $\pm$ 0.53	0.08 $\pm$ 0.48
19	Prone to mistakes/errors	0.23 $\pm$ 0.93	0.08 $\pm$ 0.43
20	Disorganised	0.21 $\pm$ 0.58	0.04 $\pm$ 0.29

Values represent means  $\pm$  SD

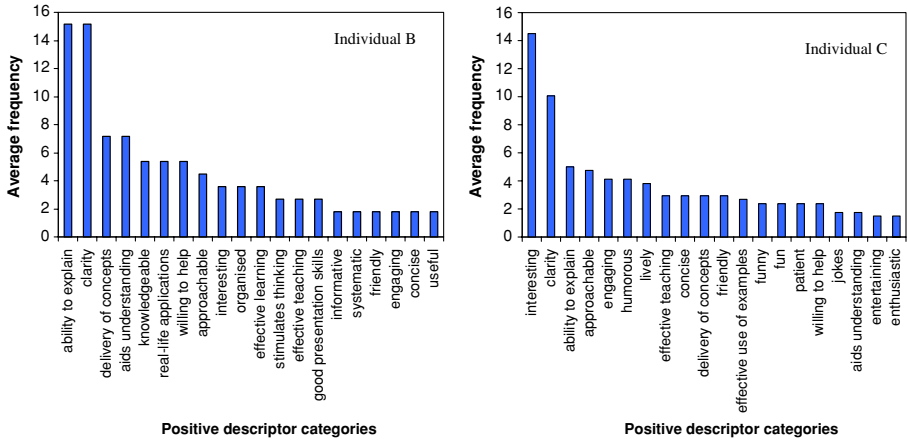
\* Significantly different from the corresponding value for the teaching award winners ( $P < 0.05$ )

Thus, while Y and Z were both given low ratings, they were for different reasons. These results indicate that the low rating obtained by individual Y was related to a lack of effectiveness in teaching, while for individual Z it was because of poor interpersonal skills and problems in dealing with students.

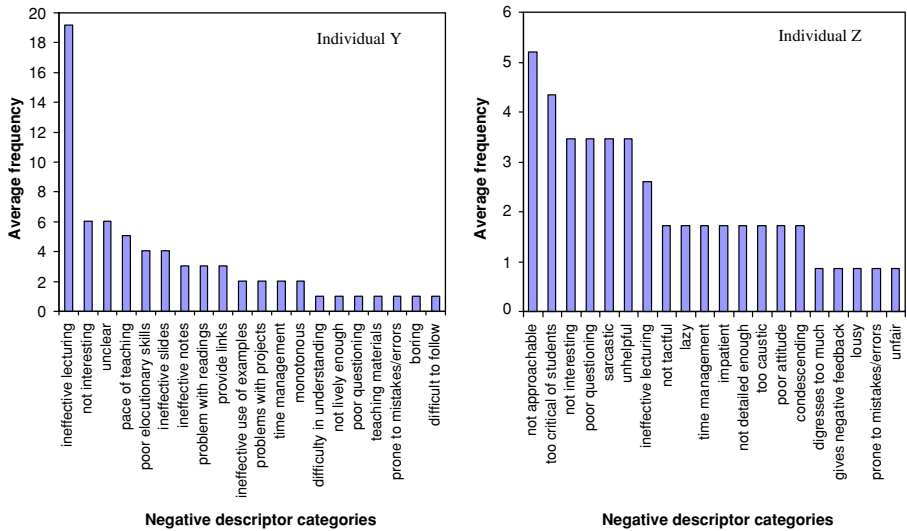
## Discussion

### Are Student Ratings Nothing more than a Popularity Contest?

At NUS, effort is constantly made to gather information that is potentially useful for decisions about assessing and improving teaching and learning. Little is known, however, about how students go about evaluating their teachers and their teaching. Hence there is some measure of cynicism. Firstly, is it a measure of student satisfaction with their teacher or the latter's effectiveness? As Abrami et al. (1990) pointed out, "Student ratings are seldom criticised as measures of student satisfaction ... [but] are often criticised as measures of instructional effectiveness". Secondly, does it really reflect the teacher's



**Fig. 5** A comparison on the personal profile of positive descriptors (frequency per 100 students) between individual B and individual C, who received a student rating of 4.557 and 4.392, respectively, for overall teaching effectiveness



**Fig. 6** A comparison on the personal profiles of negative descriptors (frequency per 100 students) between individual Y and individual Z, who received a student rating of 3.506 and 3.636, respectively, for overall teaching effectiveness

effectiveness or is it more about his/her personality, entertainment skills, and other factors not related to teaching? (Centra 1987).

In terms of bias in teaching evaluation, the issue of expected grade and grade leniency may be the most controversial of all (Marsh 2007).<sup>6</sup> Some faculty members argue that easy grading teachers get rewarded for their generosity, regardless of how well they taught (Greenwald and Gillmore 1997). Although there may be some truth to that interpretation

<sup>6</sup> See Marsh (2007) for a summary of relations between student evaluation of teaching and potential biases.

(Abrami et al. 1980; Feldman 1976; Howard and Maxwell 1980; Marsh 1987); Ory and Ryan (2001) offered another plausible explanation, namely that students who feel they have learned a great deal expect high grades for their effort and in turn rate their teachers highly for good teaching.

More important, an analysis of ratings obtained by teaching award winners at NUS reveals that, despite having an average level of difficulty similar to that of the highest 20% cohort, there is no significant correlation between rating on overall teaching effectiveness and the expected grade (Table 1). This is a novel observation and suggests that teaching award winners were able to shift students' focus momentarily away from the grade and help them focus on learning even difficult topics. They probably achieved this through teaching in "interesting" and "engaging" ways that "aid understanding", as revealed by their profile of positive descriptors.

Likewise, it has been claimed that students taking more difficult courses tend to give lower ratings. Thus, some teachers may not only assign grades more leniently, but also reduce levels of workload, on the assumption that they will be rewarded with higher evaluation ratings (Marsh 2007). However, our findings do not support this proposition. Firstly, the average level of difficulty obtained from the highest 20% cohort, lowest 20% cohort and award winners were almost identical. Secondly, there were no significant correlations between the average rating of overall teaching effectiveness and level of difficulty for all three groups of faculty members.

Another extraneous factor supposedly affecting student ratings at NUS is class size but there is evidence against this belief in the literature. Cashin (1992) argued that, taken alone, class size was not a serious source of bias. Centra (1993) also concluded that rating differences due to class size had little practical significance, and Chiu (1999) demonstrated that class size accounted for less than 1% of variance in ratings. Indeed, the significantly negative correlations obtained between the ratings of overall effectiveness and class size for the highest and lowest 20% cohort, and the larger average class size taught by lowest 20% cohort, in this study would seem to support that teachers with larger classes get lower ratings at NUS. However, it is contradicted by the fact that teaching award winners also had a significantly larger class size, but a comparable rating of overall teaching effectiveness when compared to the highest 20% cohort (Table 1). From these results, it can be deduced that outstanding teachers at NUS were able to resolve problems related to teaching large classes, and handling such classes did not handicap them. Judging by our results, it does not appear that faculty members could receive higher ratings just by teaching small classes, 'dumbing down' their course content and/or by being easy graders.

Based on the qualitative feedback, the assumption that student ratings are nothing more than a popularity contest, with the friendly, humorous, entertaining teachers emerging as winners every time, is incorrect. This is clearly evident in our findings. The profile of positive descriptors for teaching awards winners shows that attributes like "interesting", "aids understanding", and "engaging" appear at significantly higher frequency than in the general profile of an effective teacher (Table 9). It is probably due to their unique strengths in these three items that they were able to handle large classes well and to help students get over the anxiety of expecting high grades. They did not achieve this merely by "telling jokes" or being "humorous".

The findings of this study support the proposition made by Ory and Ryan (2001) that there is no strong relationship between teacher characteristics and student ratings. Some faculty members believe that students appreciate a good storyteller or a charismatic speaker, but our findings show that, as suggested by Erdle et al. (1985), if a teacher's

personality traits affect their students, it may be caused more by what they do in their teaching than by who they are. While “interesting” was ranked first in the general profile for an effective teacher at NUS, it had a high association not only with “humorous”, but also “aids understanding”, “clarity” and “ability to explain”. Equally significant is that “effective teaching” had a high association with “interesting”, “clarity”, “ability to explain” and “aids understanding”, but not with “humorous”. Thus, it becomes apparent that what really counts in an effective teacher is not who he/she is but how he/she teaches in the classroom (Ory and Ryan 2001).

Similarly, an examination of the general profile of descriptors for an ineffective teacher obtained from the lowest 20% cohort further supports the notion that students at NUS did not rate their teachers solely based on their personality or entertainment skills. The most frequent descriptor in this profile was “ineffective lecturing”, which had high shared responses with “time management”, “not interesting”, “ineffective notes”, “pace of teaching”, and “unclear”, and not with any of the other teacher characteristics (Fig. 1).

### Can Students make Meaningful Judgements?

One of the unintended consequences of using student ratings is that the university may reward poor teaching (Ory and Ryan 2001). There is a concern that students cannot make valid judgements concerning teachers and their teaching, and that teaching award winners are mostly good entertainers who receive higher ratings than less flashy but ‘better’ teachers.

However, our results suggest otherwise. As discussed earlier, teaching awards winners achieved higher frequency for items such as “interesting”, “aids understanding”, “ability to explain” and “clarity” than for “telling jokes” or being “humorous”. This shows that students’ conception of an effective teacher is more of someone engaging them intellectually rather than making teaching entertaining and fun. Showing them how concepts and knowledge can be applied in real life, asking them to grapple with new ideas, and stimulating thinking through questioning, seem to be highly valued by students (Table 9). The nature of these descriptors reflects how students can be critical and reflective observers of the teaching and learning process.

When compared with the highest 20% cohort, teaching award winners at NUS had problems mainly with “pace of teaching” and “time management”. From here, it can be deduced that they were eager to impart their knowledge to their students, resulting in weakness in these aspects.

### Going Beyond Student Ratings

It can thus be seen that using students’ written comments in a quantified manner can give a more complete picture of teaching performance, and can compensate for the limitations of solely relying on ratings. From instance, although we observed a significant difference in the ratings of overall teaching effectiveness between the highest 20% and lowest 20% cohorts, the absolute difference obtained was small. By contrast, not only were significant differences detected consistently in the frequency of almost all the positive and negative descriptors of the two cohorts, the absolute differences were substantial (Tables 3 and 4).

Moreover, using student ratings alone to discriminate between faculty members for teaching assessment cannot be regarded as ideal. Results reported here reveal that faculty members having similar scores can have totally different profiles of positive and negative



descriptors (Figs. 5 and 6). Emery et al. (2003) gave the example of how a teacher was given different scores on an evaluation questionnaire by different groups of students despite displaying exactly the same behaviour; hence he was at a loss as to how to improve his performance. This issue can be easily resolved by making use of the teacher profiles derived from the descriptors to get a deeper insight into the meaning of the ratings.

### Implications and Applications of Profiling Teacher/Teaching

With the establishment of the general profile of an effective teacher, it is possible to use it as a reference point for examining students' written comments for any individual faculty member. A graphic comparison on the strengths and weaknesses of a faculty in relation to the general profile of an effective teacher (Figs. 3 and 4) can be meaningful and revealing. This strengthens the use of student ratings as useful information for summative and formative purposes.

It is essential for faculty members to study and learn from the general profile of an ineffective teacher which, indirectly, would facilitate their understanding of students' concerns and needs. However, it is important to point out that faculty members should not look at the general profile of an effective teacher as a set of desired teaching behaviours and try to change his/her teaching style. Similarly, it should not be used by administrators to reinforce specific teaching behaviours because it would constrict rather than encourage a diversity of classroom strategies.

To counteract such unintended and undesirable consequences that could be created by the general profiles, individual faculty members must be supplied with their own "fingerprint" of positive and negative descriptors. This will enable them to know about their own unique strengths and weaknesses, and the ways in which they differ from the rest of the teaching community. Faculty members with very low ratings are often encouraged by their departments to seek help, possibly from colleagues or a campus faculty development office. As mentioned previously, it is difficult to pinpoint the remedial action to be taken by relying on ratings alone. Hence, faculty members with low ratings are often cynical about their scores or disheartened. With the establishment of the profile of negative descriptors for an individual, the most appropriate types of remedial action needed become readily apparent. By making explicit their profiles of positive descriptors, they would be strengthened psychologically to face up to negative comments and be more likely motivated to make changes and adjustments to address student concerns.

### Conclusion

Many universities, including NUS, are currently putting more emphasis on good teaching and on identifying, honouring and rewarding effective teachers. Concomitantly, attention to evaluation of teaching has intensified. Student ratings, as one of the most available sources of information, will continue to be important—if not more important—not only in providing formative feedback but also in forming personnel decisions. At NUS, much effort has been channelled into improving student feedback, particularly in deliberating and formulating the number and types of questions to be included in the quantitative survey. Practicality, however, imposes a limit on the number of items. Furthermore, even if the instrument is constructed with great care, the complaint is that a numerical score is in itself not very informative and, moreover, is subject to personal interpretations which could engender scepticism.

In this study, we demonstrated that, by using STAS, substantial information can be extracted conveniently and efficiently from students' written comments. This strengthening of quantitative data with meaningful qualitative interpretation can have important and widespread benefits and applications as listed below:

- (a) creating a profile of positive descriptors detailing what students regard as an effective teacher;
- (b) creating a profile of negative descriptors detailing what students regard as an ineffective teacher;
- (c) evaluating the strengths of an individual according to the (a) profile;
- (d) evaluating the weaknesses of an individual according to the (b) profile;
- (e) providing a personal profile (fingerprint) of strengths for an individual;
- (f) providing a personal profile (fingerprint) of weaknesses for an individual;
- (g) offering a basis for educating students about "effective teaching" in relation to "effective learning", based on (a) and (b);
- (h) enabling the use of information available from (c), (d), (e), and (f) for summative purposes;
- (i) enabling the use of information available from (b) for identifying possible corrective measures for the entire teaching community;
- (j) providing information available from (f) to enable an individual to improve his/her teaching performance;
- (k) using (a) and (b) as a baseline to enable the institution to study the developmental aspect of students' concepts of teacher and teaching over time;
- (l) documenting (c), (d), (e), and (f) to provide detailed profiles of teaching staff in the institution and allow comparison over time.

It can therefore be seen that quantified students' written comments can be presented in a meaningful way and can be readily accessible for both formative and summative applications. More specifically, the use of profiling would help identify specific areas for remedial action. In addition, the concept of profiling can be applied to other teaching evaluation tools such as the peer review report and the teaching portfolio to facilitate the quantitative interpretation of qualitative information obtained from these instruments. This will then allow for a greater integration of the various components of teaching evaluation and provide a more holistic outlook.

Student feedback continues to be relied upon at most institutions as an important teaching evaluation tool. Studies have shown that on the whole, students do provide reliable feedback. However, it is also true that using student ratings alone to discriminate between faculty members for teaching assessment cannot be regarded as complete and should therefore be augmented by qualitative interpretation of student written feedback. It is only then that we can give meaning to the numerical nature of student ratings.

## References

- Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 213–264). New York: Agathon Press.
- Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–456). New York: Springer.
- Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). The validity of student ratings of instruction: what we know and what we don't. *Journal of Educational Psychology*, 82, 219–231.

- Abrami, et al. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology*, 72, 107–118.
- Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. In L. M. Aleamoni (Ed.), *Techniques for evaluation and improving instruction, new directions for teaching and learning* (No. 31, pp. 25–31). San Francisco: Jossey-Bass.
- Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, 9, 67–84.
- Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education*, 30(6), 723–748.
- Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker Publishing Company.
- Bejar, I. I., & Doyle, K. O. (1976). The effect of prior expectations on the structure of student ratings of instruction. *Journal of Educational Measurement*, 13(2), 151–154.
- Cashin, W. E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development*, 12(2), 1–6.
- Centra, J. A. (1981). Determining faculty effectiveness. *The Journal of Higher Education*, 52(3), 328–330.
- Centra, J. A. (1987). Formative and summative evaluation: Parody or paradox? In L. M. Aleamoni (Ed.), *Techniques for evaluation and improving instruction, new directions for teaching and learning* (No. 31, pp. 47–55). San Francisco: Jossey-Bass.
- Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.
- Chiu, S. (1999). Use of the unbalanced nested ANOVA to exam factors influencing student ratings of instructional quality, unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Emery, C., Kramer, T. & Tian, R. (2003). *Return to academic standards: Challenge the student evaluation of teaching effectiveness*. Retrieved May 6, 2007, from [www.bus.lsu.edu/academics/accounting/faculty/lcrumbley/stu\\_rat\\_of\\_%20instr.htm](http://www.bus.lsu.edu/academics/accounting/faculty/lcrumbley/stu_rat_of_%20instr.htm).
- Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior and student ratings of college teaching effectiveness: A path analysis. *Journal of Educational Psychology*, 77(4), 394–407.
- Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4, 69–111.
- Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–408). New York: Agathon Press.
- Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). New York: Springer.
- Franklin, J. & Theall, M. (1989). Who reads ratings: knowledge, attitude, and practice of users of student ratings of instruction. Paper presented at the 1988 annual Meeting of the American Education Research association, San Francisco, CA.
- Gallagher, T. (2000). Embracing student evaluations of teaching: A case study. *Teaching Sociology*, 28(2), 140–147.
- Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1209–1217.
- Guthrie, E. R. (1954). *The evaluation of teaching: A progress report*. Seattle: University of Washington.
- Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, 45(3), 311–323.
- Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation. *Journal of Educational Psychology*, 72, 810–820.
- Johnson, T. & Sorenson, L. (2004). *Online student ratings of instruction, new directions for teaching and learning* (No. 97). San Francisco: Jossey Bass.
- Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education*, 40(1), 69–97.
- Lewis, K. G. (Ed.). (2001). *Techniques and strategies for interpreting student evaluations, new directions for teaching and learning* (No. 87). San Francisco: Jossey Bass.
- Lin, Y., McKeachie, W. J., & Tucker, D. G. (1984). The use of student ratings in promotion decisions. *Journal of Higher Education*, 55, 583–589.
- Lowman, J. (1996). *Characteristics of exemplary teachers, new directions for teaching and learning* (No. 65, pp. 33–50). San Francisco: Jossey-Bass.

- Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52(1), 77–95.
- Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students for different academic settings and their relationship to student/course/instructor characteristics. *Journal of Educational Psychology*, 75(1), 150–166.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, 76, 707–754.
- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253–388.
- Marsh, H. W. (1991). Multidimensional student's evaluation of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83, 285–296.
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York: Springer.
- Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education*, 64, 1–18.
- Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluation of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon Press.
- McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, 65, 384–397.
- Millman, J. (Ed.). (1981). *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- Ory, J. C. (2000). *Teaching evaluation: Past, present, and future teaching, new directions for teaching & learning* (No. 83, pp. 13–18). San Francisco: Jossey-Bass.
- Ory, J. C., & Braskamp, L. A. (1981). Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education*, 15, 271–282.
- Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. Abrami, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? New directions for institutional research* (No. 109, pp. 27–44). San Francisco: Jossey-Bass.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage Publications.
- Perry, R. P., Niemi, R. R., & Jones, K. (1974). Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. *Journal of Educational Psychology*, 66(6), 851–856.
- Remmers, H. H. (1949). Are student ratings of instructors related to their grades? Student achievement and instructor evaluation in Chemistry. *Studies in Higher Evaluation*, 66, 18–26.
- Ryan, J. M., & Harrison, P. D. (1995). The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education*, 36, 213–228.
- Theall, M. & Franklin, J. (Eds.). (1990). *Student ratings of instruction: issues for improving practice, new directions for teaching and learning* (No. 43). San Francisco: Jossey-Bass.
- Theall, M., & Franklin, J. (1999). Faculty thinking about the design and evaluation of instruction. In P. Goodyear & N. Hativa (Eds.), *Teacher thinking, beliefs, and knowledge in higher education*. The Netherlands: Kluwer Academic Publishers.
- Wirtz, J. (2004). Student feedback collection tools that can help to continuously improve your teaching. Retrieved May, 6, 2007, from <http://www.cdtl.nus.edu.sg/link/mar2004/cover.htm>.
- Wulff, D. H., & Nyquist, J. D. (2001). Using qualitative methods to generate data for instructional development. In K. G. Lewis & J. P. Lunde (Eds.), *Face to face: A sourcebook of individual consultation techniques for faculty/instructional developers* (2nd ed.). Stillwater, OK: New forums Press.