AAUSC Issues in Language Program Direction

From Thought to Action: Exploring Beliefs and Outcomes in the Foreign Language Program

H. Jay Siskin
Editor

THOMSON
HEINLE

Australia • Brazil • Canada • Mexico • Singapore • Spain
United Kingdom • United States

THOMSON
HEINLE

AAUSC 2007: From Thought to Action:
Exploring Beliefs and Outcomes in the Foreign Language Program
H. Jay Siskin, Editor

Executive Editor: Lara Ramsey Semones
Assistant Editor: Catharine Thomson
Associate Technology Project Manager: Morgen Murphy
Associate Content Project Manager: Jessica Rasile
Senior Marketing Manager: Lindsey Richardson

Senior Marketing Communications Manager: Stacey Purviance
Marketing Assistant: Denise Bousquet
Manufacturing Manager: Marcia Locke
Print Buyer: Susan Carroll
Compositor: GEX Publishing Services
Project Manager: GEX Publishing Services

Printed in the United States of America
1 2 3 — 09 08 07

Thomson Higher Education
25 Thomson Place
Boston, MA 02210-1202
USA

ISBN-13: 978-1-4282-3011-8
ISBN-10: 1-4282-3011-4

Library of Congress Control Number:
2007936134

Chapter 13
Student Evaluations and Teacher Assessment: How Much Is the Course a Reflection of the Teacher?

John D. Sundquist, Purdue University
Colleen A. Neary-Sundquist, Purdue University

Introduction

In most beginning and intermediate-level language programs with multisection courses, Student Evaluations of Teaching Effectiveness (SETEs) are a common tool by which language program directors (LPDs) gather relevant information about undergraduate students' attitudes toward aspects of foreign language study. Each semester, SETEs provide LPDs with an account of students' opinions about aspects of their teachers, course materials, and classroom activities. In light of the research on student attitudes toward foreign language study, such evaluations sometimes provide an additional source of input for understanding students' general beliefs about their language learning experience.[1] SETEs may provide relevant data for comparison with attitudinal studies and allow LPDs the opportunity to observe changing student attitudes in their programs over several semesters.

Although the types of questions and format of SETEs vary by institution and department, LPDs often use them in two ways: to assess the efficacy of a language program in general and to monitor the progress of graduate student teaching assistants (TAs). Undergraduate students are often asked to evaluate issues that pertain to their language courses, such as the choice of teaching methodology or textbook; at the same time they are asked to comment on issues that pertain to their teacher, such as his or her rapport with the class or his or her level of enthusiasm. In many large programs with multiple sections of beginning and intermediate language courses, the former issues are usually handled by the LPD and affect all sections of a course while the latter variables are largely in the hands of each individual TA and pertain only to his or her section. Thus, LPDs use the same SETEs to gather information on students' feelings about a course and their attitude toward the instructor.

This article focuses on the relationship between those two areas. In particular, we discuss the overlapping nature of students' attitudes toward their TA and their attitudes toward their language course as evidenced by responses on SETEs. By means of quantitative analysis of 80 individual German language courses in a multisection curriculum collected over six semesters at a large research institution, we provide evidence that students' perceptions of a course, regardless of the course materials or the syllabus, are often a reflection of their attitudes toward their instructor. Using findings from standard SETEs as a point of departure, we

present evidence from a follow-up survey of students in first- and second-year German courses in which we attempt to separate aspects of the course from the instructor, using a revised SETE. In the final section of the article, we discuss several implications of this study for LPDs. We suggest ways in which SETEs can be designed more effectively to reflect differences between student attitudes toward their course versus their attitudes toward their instructor. We then discuss the implications of those findings for LPDs as they consider aspects of teacher assessment and curriculum design for multisection courses in the future.

Review of the Literature

Validity and Multidimensionality of SETEs

Over the last 30 years, there has been a substantial amount of research into student evaluations of teaching effectiveness in a variety of disciplines.[2] A number of studies contain helpful overviews of the issues (Feldman, 1988; Greenwald, 1997; Harrison, Douglas, & Burdsal, 2004; Marsh, 1987). In spite of or perhaps because of the large number of studies, several areas of inquiry remain unresolved and are highly disputed. The issues most pertinent to the present study are the validity and multidimensionality of SETEs.

The discussion of the validity of students' evaluations of teaching has been characterized by a debate over the supposed multidimensionality of teaching and the ability of SETEs to reflect this accurately. Marsh and Roche (1997) argue that teaching is a complex, multidimensional activity and that evaluations therefore must consist of multiple specific items that must be chosen and worded carefully in order to be valid. For example, an effective SETE would allow students to evaluate various characteristics of the teacher separately, including his or her organization, enthusiasm, and clarity of explanation. Citing the fact that over 30 published factor analyses have provided support for this contention, Marsh and Roche follow the proposal in Marsh (1981) for the nine-factor Student Evaluation of Educational Quality (SEEQ) instrument. These nine factors are learning/value, enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examination/grading, assignments/readings, and workload/difficulty. In contrast to Marsh and Roche, d'Apollonia and Abrami (1997) argue that although teaching may be a multidimensional activity, the evidence that SETEs actually measure distinct factors is questionable. They offer an alternative analysis that indicates that student evaluation forms actually measure a single global component of overall effectiveness, which they call General Instructional Effectiveness.

Another concern about the validity of SETEs is the issue of student bias. Clayson (2005) has taken a somewhat different approach to the investigation of student bias in teaching evaluations. He points out that studies of SETEs have not taken the obvious step of asking students what kind of ratings they give to what type of instructors and why. Clayson attempted to address this issue by surveying marketing students on whether they had ever given an average teacher a higher

rating because they liked his or her personality; Clayson found that 60% of the students answered yes. Therefore, he concludes that personality and likeability can be important sources of bias in student evaluations.

Multisection courses, which are taught by different instructors but share aspects of course design, textbooks, and exams, offer a unique opportunity to explore other issues of validity and reliability. Marsh (1987) reports that the course sections with the highest SETEs were the same sections that received the highest scores on common final examinations, offering support for the validity of SETEs in predicting students' learning. In addition, in an earlier study, Marsh (1981) analyzed the relative importance of the "teacher effect" compared to the "course effect" on student ratings in multisection courses. Using data from 1,364 courses from various disciplines, Marsh (1981) determined that the effect of the teacher on student ratings is much larger than the effect of the course being taught. Furthermore, he argues that individual instructors carry with them a unique level of variance from one section of a course to another; as a result, their rating is relatively predictable when they teach the same course to a different set of students (Marsh, 1987, p. 278). These findings also corroborate those reported in Marsh and Overall (1981), where it was demonstrated that the instructor teaching a course accounts for 5 to 10 times as much variance as the course. In sum, the instructor appears to be the primary factor that shapes the outcome of SETEs of a course.

The Use of Student Evaluations in Language Programs

Much of the research on the use of standard student evaluations in language programs focuses on their usefulness in TA evaluation and supervision. As Schulz (1980) points out, although many language departments did not implement SETEs until the 1960s and 1970s, by 1979, over 80% of the 196 language departments in Schulz's survey made use of either university-wide evaluations or departmentally designed evaluations (p. 4). It is safe to assume that the trend has continued and that to varying degrees, most LPDs rely on SETEs to assess TAs' classroom performance. Early discussion of teacher preparation in language programs has shown that student evaluations played an important role for both LPDs and TAs. Di Donato (1983) and Herschensohn (1992) list midterm or end-of-term student evaluations as one of the recommended essentials in TA training, citing their importance as a means to provide inexperienced TAs with feedback.

Most recently, Brandl (2000) notes that TAs believe that SETEs are one of the most important mechanisms by which they might improve their teaching. Along with informal discussion with a supervisor or peers, student evaluations were seen as a crucial but low anxiety-causing training component (p. 359). Brandl (2000) points out that TAs unanimously agree that SETEs should be taken seriously and that many of them consider "the students' ratings like a course grade that reflected the quality of their teaching" (p. 364). However, others such as Brinko (1993) warn that when SETEs are used as a source of feedback on instruction, they must be used by both TAs and supervisors in conjunction with other types of assessment. Barnett (1983) and Magnan (1993) echo this sentiment by suggesting that end-of-course student evaluations be used along with such other training tools as classroom observations by LPDs, videotaped class sessions with peers or a supervisor, and peer observations and informal discussions.

Although SETEs are most commonly used for TA training, they often serve an additional purpose as a means to evaluate the language program curriculum in general. In particular, open-ended questions on many evaluation forms allow students the freedom to express their enthusiasm or concerns regarding various aspects of the course, textbook materials, or grading. As Lynch (1996) discusses throughout his book on language program evaluation, student evaluations are one of several useful tools for gathering quantitative and qualitative data on the efficacy of a language program. Brown (1995) makes the same suggestion, pointing out, however, that SETEs often address too many issues at once (p. 201). He notes that LPDs must make sure that evaluations are clear and concise and that if student evaluations address the quality of the language curriculum in addition to the quality of a teacher's instruction, these issues should be clearly separated (p. 201). Brown (1995) emphasizes the danger of using SETEs to evaluate the efficacy of a language program since administrators often use only the numerical results of the evaluations rather than student responses to open-ended questions. As was clear in the earlier discussion on teacher training, SETEs are most useful when used as a supplement to other tools of curriculum evaluation.

The Study

Background

The present study focuses on SETEs from the first four multisection courses in the beginning and intermediate-level undergraduate German language sequence at a large research institution in the Midwestern United States. Each section of these courses is taught by a graduate TA who is responsible for the day-to-day activities of the class, including in-class instruction, lesson planning, assigning and grading of homework, and grading of tests. All four courses are overseen by the LPD, who selects materials; designs the syllabus and course schedule; writes the tests; and establishes grading criteria for homework, tests, and final grade weighting. Although the numbers vary each year in this program, there are approximately 12 native-speaker and non-native speaker TAs in total with varying levels of experience, who teach about 18 sections of the four courses each semester with an average of 25 students in each section. Students in each section have a wide variety of majors and minors. Most students note on the evaluations that they are taking the course to fulfill the requirements of their degree rather than taking it as an elective.[3]

The SETEs are administered the last week of each semester. The evaluations are anonymous and are not returned to the instructor until several months after the undergraduate students' final grades have been turned in.[4] The standard SETE was developed by the university's language department to be used by instructors of all languages and includes a series of Likert-scale questions and four open-ended questions with space for student comments. Of the 18 Likert-scale questions, 16 pertain to specific aspects of the instructor's performance, while the last 2 items ask students to evaluate overall effectiveness of the course and overall effectiveness of the instructor. The open-ended questions elicit comments from students on both the instructor and the course.

We gathered data from SETEs collected over six different semesters between Fall 2002 and Spring 2006. Focusing on the computer-generated summative data sheet from each section's SETEs, we analyzed a total of 80 sections of German language courses.[5] The total number of students who participated in the evaluations over the six semesters is 1,424, and the total number of different TAs evaluated by the students is 30.

Data Analysis

Measurements of Overall Effectiveness of the Course and the Instructor

From each TA's summative sheet of his or her students' evaluations, we gathered information on the two questions about global performance:

1. Overall, I would rate this **course** as Excellent, Good, Fair, Poor, Very Poor.
2. Overall, I would rate this **instructor** as Excellent, Good, Fair, Poor, Very Poor.

   (Excellent = 5, Good = 4, Fair = 3, Poor = 2, Very Poor = 1)

Those two items were of interest to us because they are designed to evaluate overall effectiveness and they provide students with the opportunity to assess the instructor and the course separately.[6]

The median scores on these two items were then used to determine the extent to which there is a correlation between them. To demonstrate the way in which the data were collected, we have included figures from a sample semester. Listed in Table 1 are the median ratings for the two items for each section, along with a list of the instructors and the number of students in each section during this semester.

Table 1
Semester results of questions 1 and 2 from summative sheets of SETEs

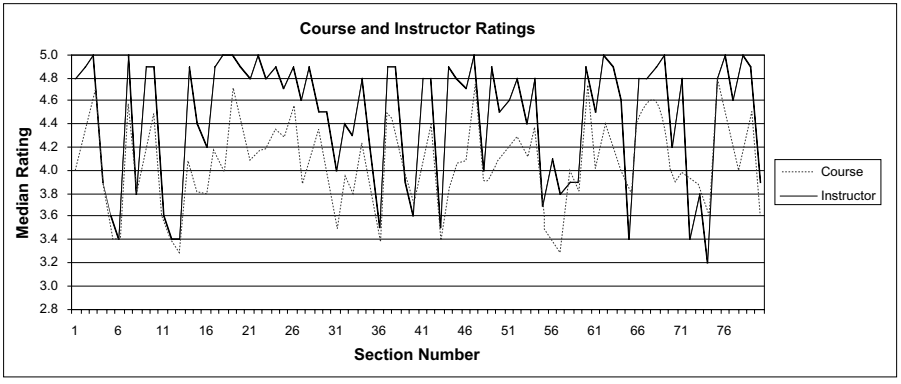
Section number	Instructor number	Course	Median rating on item 1	Median rating on item 2	Number of SETEs submitted
Section 21	Instructor 11	German 101	4.2	4.9	13
Section 22	Instructor 11	German 101	4.7	5.0	22
Section 23	Instructor 12	German 101	4.4	4.5	19
Section 24	Instructor 13	German 101	4.0	4.4	21
Section 25	Instructor 7	German 102	4.1	4.8	15
Section 26	Instructor 4	German 102	3.9	4.6	13
Section 27	Instructor 4	German 102	4.0	4.9	12
Section 28	Instructor 6	German 102	4.5	4.9	21
Section 29	Instructor 6	German 102	4.2	4.9	18
Section 30	Instructor 10	German 201	4.2	5.0	22

Table 1 (continued)
Semester results of questions 1 and 2 from summative sheets of SETEs

Section number	Instructor number	Course	Median rating on item 1	Median rating on item 2	Number of SETEs submitted
Section 31	Instructor 10	German 201	4.2	4.8	24
Section 32	Instructor 3	German 201	3.6	4.2	20
Section 33	Instructor 3	German 201	3.4	3.5	16
Section 34	Instructor 9	German 201	4.3	4.7	15
Section 35	Instructor 9	German 201	4.6	4.9	23
Section 36	Instructor 1	German 202	4.0	4.5	15
Section 37	Instructor 1	German 202	3.5	4.0	20
Section 38	Instructor 8	German 202	4.4	4.9	17
Section 39	Instructor 5	German 202	3.9	3.9	17

The results from this single semester are indicative of trends in the SETE data from all six semesters of the study. First, the median rating for item 1 regarding the evaluation of the course is consistently lower than the median rating for item 2, the rating of the instructor. In other words, students generally rate their instructor higher than the course. Second, there is much variation across sections in terms of both the course ratings and the instructors' ratings. The range of scores for item 1 is from 3.3 to 4.8 (mean=4.1, median=4.0, standard deviation=.39) while the range for instructor ratings from item 2 is 3.2 to 5.0 (mean=4.4, median=4.8, SD=0.55). Third, and more importantly for the current study, the higher median ratings for the instructor generally correspond to higher ratings for the course. Conversely, a lower rating for an instructor corresponds to a lower rating for the course. This trend can be tracked when we view the data from all 80 sections of classes, as shown in Figure 1.

Figure 1
Course and instructor ratings



From this graph, it is clear that the two lines rise and fall in tandem and reflect a correlation. The SETEs for Section 12, for instance, reflect a high course rating of 4.5 and a high instructor rating of 4.9, while the median scores for Section 13 include a low course rating of 3.6 and an equally low instructor rating of 3.4. As we saw for the sample semester, students who rate an instructor favorably usually rate the course favorably as well and those who have a less favorable impression of the instructor rate the course accordingly. A two-tailed Pearson correlation test confirms this conclusion. There is a very high positive *r*-value for correlation between the median ratings on items 1 and 2 for all 80 sections: The *r*-value is .80, and this correlation is significant at the .01 level.

For almost all of the sections, the instructor rating is slightly higher than the course rating. The average difference between the two ratings for all of the sections is +0.42. However, there are two anomalous patterns. In 6 of the 80 sections, the instructor rating is lower than course rating. For instance, in Sections 66 and 77, students rated the instructors at 3.4 but gave the courses an overall rating of 3.8 and 3.9, respectively. Also, in some cases, especially those that involve a highly experienced TA, the difference between the two scores is much greater than the average of +0.42. Sections 19 or 45, for example, have instructor ratings of 5.0 and 4.9 with course ratings of only 4.0. Other than those two exceptional patterns, however, the course and instructor ratings correspond closely for all 80 sections analyzed.

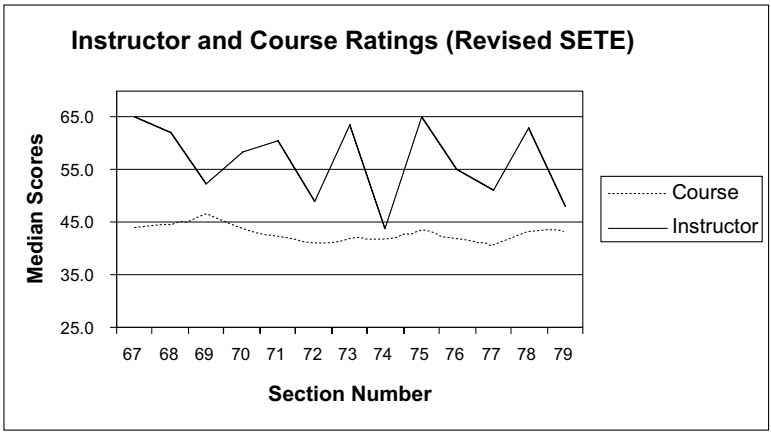
To summarize: Data from median scores of two Likert-scale global performance indicators on SETEs indicate a significantly positive correlation between students' attitudes toward their course and their attitudes toward their instructor. Although there are some sections in which students rate their instructor lower than the course, the general trend is consistent throughout the dataset.

Multidimensionality and a Revised SETE

As a follow-up survey, we administered a revised SETE form to students in the same Beginning/Intermediate German language classes during the final semester of this study. This new evaluation form, included in Appendix A, consists of 24 Likert-scale questions. Unlike the standard SETE, however, this evaluation has two clearly marked, separate sections of survey items: The first 11 items focus solely on aspects of the course, while the last 13 items relate specifically to issues that involve the instructor. The two sections are clearly marked with the titles "Course" and "Instructor" and are separated by a page turn. Students were given this evaluation before they filled out the standard SETE and were told that it was optional.[7]

To arrive at a set of figures from the new SETE that could be used for comparison with the data from the standard SETE, we calculated scores for the "Course" and "Instructor" sections separately. Using the same scale as before (5 = Strongly Agree, 4 = Agree, 3 = Fair, 2 = Disagree, and 1 = Strongly Disagree), we calculated the median rating from the total scores on the 11 items from the "Course" section of each TA's SETEs and then followed the same procedure for the 13 items from the "Instructor" section.[8] Those scores for the semester are provided in Figure 2.

Figure 2
Instructor and course ratings (revised SETE)



As is clear from the graph, the revised SETE shows general agreement among students across sections on aspects of their courses, but widespread variation on items pertaining to the instructors. This pattern differs from that of the standard SETEs in which instructor and course ratings move in tandem with each other.

Although looking at the median scores clearly shows the pattern of course versus instructor ratings allowing us to compare median scores on both the standard and the revised SETE, it also can be illuminating to look at the revised SETE data in terms of mean scores. This allows us to compare scores on a 5-point scale (median scores for each section from the revised SETEs can be found in the Appendix). We calculated two separate mean ratings by dividing the median score by the number of items in each section of the evaluation. The ratings for each instructor and course are listed in Table 2.

Table 2
Means for "course" and "instructor" questions on revised SETE

Section number	Instructor number	Course	Median rating for "course" items	Median rating for "instructor" items	Number of students
Section 67	Instructor 5	German 202	4.0	4.8	17
Section 68	Instructor 24	German 102	4.0	5.0	15
Section 69	Instructor 19	German 201	4.0	4.8	12
Section 70	Instructor 13	German 202	3.8	4.2	16
Section 71	Instructor 3	German 202	3.9	4.6	20

Table 2 (continued)
Means for "course" and "instructor" questions on revised SETE

Section number	Instructor number	Course	Median rating for "course" items	Median rating for "instructor" items	Number of students
Section 72	Instructor 29	German 101	3.7	3.8	11
Section 73	Instructor 29	German 101	4.0	3.7	13
Section 74	Instructor 22	German 202	3.8	3.4	21
Section 75	Instructor 30	German 201	4.0	5.0	18
Section 76	Instructor 30	German 201	4.1	4.9	15
Section 77	Instructor 2	German 202	3.9	4.5	14
Section 78	Instructor 26	German 102	3.8	4.9	22
Section 79	Instructor 28	German 101	4.2	4.0	18
Section 80	Instructor 28	German 201	3.7	3.9	15

The figures from the revised SETE indicate similarities with the standard SETE, although there are also some noticeable differences. One similarity is that students again rate their instructor higher than the course. For the most part, students respond more favorably on items that pertain to the instructor than on those that pertain to the course. Only sections 73, 74, and 79 are exceptions to that rule. Another similarity with the standard SETE is that the ratings for instructors on the revised SETE vary widely. Instructor ratings range from as low as 3.4 to as high as 5.0 (range=1.6, mean=4.39, SD=0.55). A major difference, however, is that there is relatively little variation among ratings from the "Course" section of this evaluation form. The lowest rating is 3.7 and the highest is 4.2 (range=0.5, mean=3.92, SD=0.15).

Another major difference between the results from the revised SETE and the standard SETE is that there is a nonsignificant correlation between the course and instructor ratings. Recall that there was a highly significant positive correlation between the two ratings (r=.80) on the standard SETEs. However, a two-tailed Pearson correlation test of the median scores from the "Course" items and "Instructor" items on the revised SETE indicates that there is a weaker correlation between the two items (r=.34) and that this correlation is not significant ($p>.05$). We can attribute this statistical pattern to the relative consistency of course ratings on the revised SETE: As we saw from the figures in Table 2, the ratings from the course items do not fluctuate very much and do not rise and fall in tandem with the instructor ratings as much as they did on the standard SETE. Thus, it appears that the revised SETEs allow the students to evaluate more aspects of the instructor without transferring their attitudes to their evaluation of the course.

Discussion

Student Attitudes and the Standard SETE

The data in the previous section indicate that students' attitudes toward their immediate language learning experience in the classroom closely reflect their attitudes toward their instructors. These findings corroborate those in Marsh (1981, 1987) where it was determined that the instructor is the primary factor in evaluations of multisection courses. In the first part of the empirical analysis, we examined global performance indicators that revealed the overlapping nature of student attitudes in these two areas. When asked to evaluate the overall effectiveness of a course, students rarely deviate from their own opinion of the instructor. When they believe that an instructor is effective, they also approve of the course, including its materials, syllabus, and grading. On the other hand, if they believe that the instructor is not effective, they evaluate the course negatively, despite identical course content and organization. The data on the two general performance items in the previous section reveal a strong correlation across sections, courses, and levels.

There are several possible reasons for this pattern. The first is purely practical and relates to the way in which the standard SETE is structured. Since all 16 of the individual questions on this standard SETE relate to the characteristics of the instructor, this may prime the students to allow their opinion of the instructor to influence their opinion of the course. That is, they are not forced by the standard SETE to think about the characteristics of the course in any way until they are asked to give a global rating of its effectiveness. Therefore, they may be likely to give the course an overall rating that is similar to that of the instructor.

A second possible reason for the correlation between course and instructor ratings involves the students' lack of understanding of how multisection courses are developed. They may not realize that the course is designed (at least at this university) by the LPD, who chooses the textbook, determines the percentages for the grade breakdown, and writes all exams. Instead, they may assume that the instructor has more control over the design of the course, as is likely the case in many of their other classes. If the students do not realize that the course is designed by one person and taught by another, who may have slightly different views on teaching, they are not likely to separate the course and the instructor in their minds and consequently are not able to rate them separately.

Another reason the global performance indicators are so strongly correlated relates to the particular nature of communicative language teaching. In a communicative classroom, interaction between the students as well as between students and the teacher is foregrounded. Thus, the teacher has a greater presence in this type of classroom as opposed to a class with a lecture format. In a traditional lecture course format, for example, it might be easier for students to separate the instructor from the material he or she is lecturing on because they might perceive the instructor to be primarily "delivering" the course content rather than using it as a means to interact with them. In a communicatively oriented classroom in which the teacher is an active participant, the teacher's presence may be so

sufficiently noticeable to the students that they associate the teacher with the course more strongly. They would transfer any attitudes they might have about the instructor toward the course. If as Clayson (2005) notes, many students' perceptions of a course are largely dependent on aspects of the instructor's personality, this effect might be even greater in an interactive classroom environment.

A final explanation for the high correlation between course and instructor rating is that the students are quite accurate in associating the course with the instructor so closely. Although the LPD designs the course, it is the instructors who teach it; and their teaching can substantially alter the course in spite of the materials the LPD has designed. For example, if students are asked whether the tests in a course are fair, what factors would be relevant in their answer? A student would probably answer that a test is fair when the material on the test was covered in class in such a way that they knew what was important. The test would also be fair only if the teacher made sure that the students understood the material and explained anything that was unclear. The way in which a test is graded can also make it "unfair." In the same way, if students were asked whether the textbook is helpful or interesting, their answer would most likely depend not on their objective evaluation of the materials, but on their experience with how the instructor used the textbook in class—from the selection of particular activities to the way those activities were conducted. The same textbook could seem very different in the hands of different instructors.

From those two examples, it is clear that any test that the LPD writes can be made fair or unfair to students by what the teacher does or does not do in class on a daily basis. Therefore, if students give the class and the teacher a similar ranking, it may simply reveal something about the effectiveness of the teacher and the way he or she manipulates the course materials provided by the LPD.

The Revised SETE

The revised SETE, which asked 11 individual questions about the course and 13 about the instructor, showed a considerably smaller correlation between course ratings and instructor ratings. This suggests that an evaluation with specific questions addresses some of the potential problems with the global ratings used in the standard SETE. Asking specific questions about the course and the instructor alleviates the problem that a list of questions about the instructor might influence the students to think primarily about the instructor when they are asked to give a global rating for the course. Asking a number of specific questions about different elements of the course also may allow the students to gain some insight into thinking about the course separately from the instructor and then to evaluate each one separately.

This finding also offers support for Marsh and Roche (1997), who contend that global evaluations of teaching effectiveness are not valid because of the multidimensional nature of the activity of teaching. The global performance indicators on the standard SETE here provide relatively little reliable, unbiased information to give the LPD an idea of the progress of a TA or the effectiveness of a course. The separate, individual dimensions of the teacher and the course addressed on the revised SETE, however, provide relevant, useful information on students' attitudes about the course and instructor without combining the two.

According to the results here, the "teacher effect" seems to be quite noticeable in multisection courses in which the instructor is not responsible for the overall design and organization of the course. In this unique situation, it is more appropriate to view the multidimensionality of teaching on two separate planes: multidimensional aspects of the instructor and multidimensional aspects of the course. Therefore, in terms of Marsh's (1981) SEEQ instrument and its nine separate factors of teaching, the results of the revised SETE indicate that these two "dimensions" of multidimensionality should be kept separate from each other.

Conclusion: Implications for the LPD

The results of this study offer three important insights for the LPD to consider. The first is that it is critically important to consider the current design of a program's SETEs, given our results that standard SETEs often conflate the students' ratings of the instructor and the course. Therefore, the LPD should look closely at the SETEs provided by the university, considering the order, structure, and degree to which items pertain to the course or the instructor. LPDs may want to look at the correlations between course and instructor in the existing evaluations. In other words, the SETE is a valuable but limited tool for language program evaluation, and the LPD must be aware of its limitations.

The second interesting outcome of these results is relevant to the LPD who can invest the time and effort to redesign the SETE in order to monitor the effectiveness of changes to either curriculum design or teacher training. The example of the revised SETE here provided more precise results that allowed the LPD to separate these two aspects of the program, but there is still much room for improvement. Although a weaker correlation exists between course and instructor ratings on the revised evaluation, there is still a noticeable teacher effect on course attitudes. Future research may indicate that it is not possible to separate these two entirely in multisection language programs. Additional research also may provide insight into better types of questions or a greater variety of dimensions that should be covered by student evaluations in multisection language programs. A possible model may be the nine-factor SEEQ proposed by Marsh (1981). Keeping the multidimensionality of teaching and course separate, LPDs may be able to cater this general model to fit the needs of a foreign language department.

A final implication of this study is that SETEs are only one of a number of tools at the LPD's disposal that can and should be used to evaluate and promote effective teaching. It is important not to rely solely on any type of SETE for information about what is going on in the classroom or to expect the results of SETEs to do the work of TA training or program development. A number of other strategies (such as informal discussions with peers or supervisors, classroom observations and follow-ups, and mentoring programs) are necessary to provide TAs with the resources they need to improve their teaching and to give the LPD insight into improving the program. SETEs are a crucial part of this process, but they must be used in conjunction with other tools of measurement and evaluation.

Notes

1. For general discussion of student attitudes toward foreign language learning, see Gardner and Lambert (1972), Horwitz (1988, 1989), and Bacon and Finnemann (1990). For a helpful overview and more recent general analysis of students' beliefs about language learning, see Kern (1995), Cotterall (1999), Rifkin (2000), and Graham (2006).
2. It should be noted that very few of these studies focus on foreign language teaching.
3. At this particular institution, there is a four-semester language requirement for students in liberal arts and education and a three-semester language requirement for science majors. Other disciplines, including Engineering, Pharmacy, and Business, do not have a foreign language requirement.
4. TAs distribute the evaluations to students in class and are asked to leave the room while students fill out the forms. A student representative is asked to take the completed evaluations directly to the department's administrative staff in charge of processing them.
5. For each section of the courses, a summative data sheet of the SETEs displays the students' median rating from 0.0 to 5.0 on each of the 18 Likert-scale items.
6. Note that item 2 is the only Likert-scale question on the evaluation that specifically refers to the course and not the instructor.
7. Like the standard SETE, the revised SETE was anonymous and the results were not reported to instructors until several months after the final grades were submitted.
8. The maximum score for the first set of questions is 55 (11 questions x 5 [for "strongly agree"]), while the maximum score for the second set is 65 (13 questions x 5).

References

Bacon, S., & Finnemann, M. (1990). A study of attitudes, motives, and strategies of university foreign language students and their disposition to authentic oral and written input. *Modern Language Journal*, *74*, 459–473.

Barnett, M.A. (1983). Peer observations and analysis: Improving teaching and training TAs. *ADFL Bulletin*, *15*, 30–33.

Brandl, K. K. (2000). Foreign language TAs' perceptions of training components: Do we know how they like to be trained? *Modern Language Journal*, *84*, 355–371.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching. *Journal of Higher Education*, *64*, 574–593.

Brown, J. D. (1995). *The elements of language curriculum*. Boston: Heinle & Heinle.

Clayson, D. (2005). Within-class variability in student-teacher evaluations: Examples and problems. *Decision Sciences Journal of Innovative Education, 3*, 109–124.

Cotterall, S. (1999). Key variables in language learning: What do learners believe about them? *System*, *27*, 493–513.

D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *Journal of Educational Psychology*, *52*, 1198–1208.

Di Donato, R. (1983). TA training and supervision: A checklist for an effective program. *ADFL Bulletin*, *15*, 34–36.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty view: Matched or mismatched priorities? *Research in Higher Education*, *28*, 291–344.

Gardner, R., & Lambert, W. (1972). *Attitudes and motivation in second language learning*. Rowley, MA: Newbury.

Graham, S. (2006). A study of students' metacognitive beliefs about foreign language study and their impact on learning. *Foreign Language Annals*, *39*, 296–309.

Greenwald, A. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*, 1182–1186.

Harrison, P., Douglas, D. K., & Burdsal, C.A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*, 311–323.

Herschensohn, J. (1992). Teaching assistant development: A case study. In J. C. Walz (Ed.), *Development and supervision of teaching assistants in foreign languages* (pp. 25–46). Boston: Heinle & Heinle.

Horwitz, E. (1988). The beliefs about language learning of beginning university foreign language students. *Modern Language Journal*, *72*, 283–294.

Horwitz, E. (1989). Facing the blackboard: Student perceptions of language learning and the language classroom. *ADFL Bulletin, 20*, 61–64.

Kern, R. G. (1995). Students' and teachers' beliefs about language learning. *Foreign Language Annals*, *28*, 71–92.

Lynch, B. K. (1996). *Language program evaluation*. Cambridge: Cambridge University Press.

Magnan, S. S. (1993). Assigning new TAs to second-year courses. *ADFL Bulletin, 24*, 36–43.

Marsh, H. (1981). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*, 47–60.

Marsh, H. (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., & Overall, J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. *American Educational Research Journal*, *18*, 103–112.

Marsh, H., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, *52*, 1187–1197.

Rifkin, B. (2000). Revisiting beliefs about foreign language learning. *Foreign Language Annals, 33*, 394–420.

Schulz, R. (1980). TA training, supervision, and evaluation: Report of a survey. *ADFL Bulletin, 12*, 1–8.

Appendix A

Revised Student Evaluation of Teaching Effectiveness

The following survey of your German course and instructor is optional and should take you only about two minutes. The survey is completely anonymous. Thank you for your input—it is greatly appreciated!

Please read each statement carefully; then select one of these five alternatives:

Strongly Agree (SA), Agree (A), Undecided (U), Disagree (D), Strongly Disagree (SD).

Course

1. The textbook for this course is helpful.
2. The workbook's listening comprehension activities are useful.
3. The pace of this course is fine.
4. The number and type of projects are appropriate.
5. The different texts that we read in German enabled me to learn.
6. The tests for this course are fair.
7. The amount of homework assigned is fair.
8. The attendance policy is fair.
9. The percentage for each component of the final grade is appropriate.
10. The size of the class/number of students is good for learning a language.
11. I think this course is well-designed.

Instructor

1. My instructor is a likeable person.
2. My instructor is organized.
3. My instructor is well prepared for each lesson.
4. My instructor speaks an appropriate amount of German.
5. The way my instructor grades tests and projects is fair.
6. My instructor speaks an appropriate amount of English.
7. I am engaged in learning during class sessions.
8. My instructor creates an atmosphere beneficial to learning.
9. My instructor designs interesting activities for class sessions.
10. My instructor plans a variety of activities for class sessions.

11. My instructor is knowledgeable about German-speaking culture.
12. My instructor is knowledgeable about the German language.
13. My instructor makes me want to learn more about German culture.

Appendix B

Results of the Revised SETE (with median ratings)

Section number	Instructor number	Course	Median rating for "course" questions	Median rating for "instructor" questions	Number of students
Section 67	Instructor 5	German 202	44	63	17
Section 68	Instructor 24	German 102	44	65	15
Section 69	Instructor 19	German 201	44.5	62	12
Section 70	Instructor 13	German 202	42	55	16
Section 71	Instructor 3	German 202	42.5	60.5	20
Section 72	Instructor 29	German 101	41	49	11
Section 73	Instructor 29	German 101	44	48	13
Section 74	Instructor 22	German 202	42	44	21
Section 75	Instructor 30	German 201	44	65	18
Section 76	Instructor 30	German 201	45	64	15
Section 77	Instructor 2	German 202	43.5	58.5	14
Section 78	Instructor 26	German 102	42	63.5	22
Section 79	Instructor 28	German 101	46	52.5	18
Section 80	Instructor 28	German 201	41	51	15