MULTIPLE PERSPECTIVES ON TEACHER EVALUATION IN THE FOREIGN LANGUAGE CLASSROOM

Sara Beaudrie
Alan Brown
Greg Thompson
University of Arizona

Defining what makes up teaching effectiveness has proved to be a complex task for researchers in language pedagogy. The present study attempts to shed light on the perceptual differences of effective teaching by providing a comparison of the evaluations of teaching effectiveness of one instructor with those of his/her students from two beginning Spanish classes and three peer evaluators. Furthermore, this study provides insight into the factors that students consider when filling out university Teacher-Course Evaluation Forms (TCE). In the study, the students completed a five-item questionnaire from the TCE forms where they evaluated their instructor's teaching effectiveness and explained what factors they had taken into account in rating the instructor. Next, all the participants completed a 35-item questionnaire evaluating the instructor's effectiveness. The researchers found a significant difference between all of the participants except between the observers and one of the classes. Given these results, this research supports the notion of using multiple perspectives in teacher evaluation. In addition, this study raises some concerns regarding the validity of student and self-evaluations, hence the TCE may not truthfully reflect the teacher's effectiveness. The qualitative data showed a wide variety of reasons for students' responses that did not always correlate to the numerical score given the instructor.

INTRODUCTION

A myriad of questions arise when one begins to search for a definitive description of what makes up effective foreign language (FL) teaching and how that translates into daily concrete pedagogical practice in the classroom—not the least of which is "What does 'right' mean?". The picture blurs even more when one considers the ever-changing nature of FL and second language (SL) theories and pedagogy. Language learning theories frequently rise and fall from favor along with their accompanying methodologies which complexifies the issue of what effective language teaching may, or may not, entail. Mitchell and Vidal (2001) liken the dynamic nature of language learning theories and methods to the ebbs and flows of a river, arguing against the incomplete analogy of a pendulum. However, the pendulum metaphor does help visually represent the cyclical nature of second language acquisition (SLA) theory and pedagogy. Thus, the definition of effective language teaching appears to be dynamic as well as cyclical and iterative. This complex dynamic is further complicated by the multiple viewpoints different stakeholders bring to the enterprise of teaching and learning, i.e. teachers, students, administrators, supervisors, and researchers.

Clearly not all of those interested in SL and FL pedagogy share the same notion of 'effective' teaching. Formally trained teachers with background in SLA theory might hold up one standard while their beginning students and their supervisors hold up another. It is the intersection of these perceptions and standards that has serious consequences for language learning and teaching with the potential to cause grave misunderstanding and disillusionment or marked growth and gratification in the language classroom (Horwitz, 1990).

The desire to better understand this intersection between differing notions of effective teaching provided the impetus behind the current research. If teachers, students, and supervisors can better understand each other's perspectives regarding effective teaching, then positive gains can be made in the field of language pedagogy. Therefore, in order to achieve an increased understanding of effective language teaching and how it is perceived from different perspectives, this paper will be divided into three sections. First, a review of relevant literature from the field will be presented which addresses previous work into efforts to define effective teaching vis-à-vis students', peers', and teachers' own evaluations of teaching practices. Second, a detailed explanation of the methods and procedures of the empirical study will be presented along with results and data analysis. Finally, a discussion of the results and the implications for the field, both pedagogically and theoretically, will be included.

While the literature is replete with studies delving into the perceptions of learners and teachers concerning different aspects of teaching and learning (Brosh, 1996; Reber, 2001; Schulz, 1996; Wennerstrom & Heiser, 1992) relatively few have specifically compared and contrasted the perceptions of individual teachers regarding their teaching with those of their own students (Moore, 1996). That is to say that students' and teachers' perspectives on the exact same object of observation remains largely unattended, i.e. the same specific language class over a semester given by the same teacher. Furthermore, as far as the authors are aware, no study in FL and SL language learning has provided an additional quantitative comparison of student and teacher perceptions with those of a third-party such as peer evaluators. The present study fills this gap in the literature by providing a comparison of the perceptions of one teacher's language pedagogy with those of the students' and three peer observers. Therefore, the following research questions have been drafted to focus the current study:

> 1-How do students' evaluations of specific language teaching practices coincide with or differ from those of their teachers' when assessed using the same evaluation instrument?

2-How do these student and teacher evaluations of language teaching coincide or differ from those of peer evaluators using the same instrument?
3-Do student evaluations of language teaching on a language-specific instrument coincide with their evaluations on a more global, university-wide evaluation form?
4-What do students take into account when completing the university's TCE form?

REVIEW OF LITERATURE

In investigating current thinking on what constitutes effective FL and SL language teaching three sources have come to the forefront: 1) national standards carefully drafted by two large professional teaching organizations—ACTFL and TESOL, 2) assessment instruments used by teaching supervisors and trainers from around the United States, and 3) actual survey and questionnaire research done with teachers and learners.

The American Council on the Teaching of Foreign Languages (ACTFL) and the association of Teachers of English to Speakers of Other Languages (TESOL) represent two large professional language teaching organizations that generate standards of effective teaching for which their members may aspire. These standards, although general, have been carefully articulated and include central concerns in the teaching of languages. The ACTFL standards are comprised of five general categories which shape their prescribed standards of effective teaching, namely, communication, cultures, connections, comparisons, and communities. They reflect a desire to take the language outside of the classroom by not only addressing linguistic and pedagogical issues, but rather cultural, societal and interdisciplinary concerns.
In contrast to the ACTFL standards, which apply to all levels of instruction, TESOL has drafted several sets of standards according to level. The standards for P-12 teacher education programs incorporate 5 domains with a total of 13 overarching standards. Domains 1-4 relate directly to the interaction between teacher and student; they are language, culture, planning, implementing and managing instruction, and assessment. The standards do not prescribe specific exercises or activities for the classroom but they do provide a lengthy description of what effective ESL teachers do and what attributes they possess. The mere length of the standards espoused by such organizations as ACTFL and TESOL can be intimidating and rather unrealistic for both non-native *and* native-speaking teachers of a language.

A more realistic reflection of how standards of effective teaching reach the classroom and are concretely applied comes in the form of evaluation instruments used by teacher trainers. The Virginia Beach City Public Schools (2000) uses an evaluation form divided into five main categories: target language use, classroom activities, instructional strategies, classroom environment, assessment. The Department of Foreign Languages from the University of Arkansas has used an assessment instrument comprised of two parts: the instructor and the students. The instructor section contains 11 statements such as "command of target language," "active student involvement", etc which are rated on a 5-point scale. The descriptors used in these evaluation instruments, such as "communicative interaction in the target language (personalization of vocabulary and structures learned . . .)," make certain assumptions about what effective language teaching is. Most classroom language teachers, in whichever context and at whichever level, are held more to the standards that appear on their supervisor's evaluation form than they are to national standards.

One of the most salient studies into the perceptions language teachers have of effective teaching is provided by Reber (2001) in her dissertation research with post-secondary FL teachers throughout the western part of the United States. An 80-item questionnaire was mailed to 1,000 post-secondary instructors, 950 .of whom were members of ACTFL. After an extensive literature review, Reber personally developed the instrument and distributed the 80 items over nine different categories: 1) ACTFL's *Standards for FL Learning*, 2) corrective feedback; 3) theories and teacher behaviors related to communicative approaches; 4) focus on form in classroom SLA; 5) individual learner differences in FL learning; 6) strategies for FL learning; 7) theories about SLA; 8) teacher qualifications; 9) assessment in FL teaching. These categories gave shape to each individual item included in the questionnaire.

The 457 FL teachers who responded to the questionnaire agreed with the majority of items related to the ACTFL *Standards*, communicative language teaching, small group work, and strategies for FL learning. However, only one of the eight items pertaining to the teaching of grammar and two of the eight items relative to assessment in FL teaching reflected a high level of agreement. Furthermore, 14 of the items on the questionnaire, almost 20%, did not receive a majority agreement or disagreement. The author proposes that this response pattern may be indicative of controversial areas in SLA and FL teaching, such as error correction, Krashen's Monitor Model, and assessment. As evidenced by Reber's research, certain fires within SLA theory and pedagogy have yet to be put out and continue to smolder.

Brosh (1996) also collected data on the perceived characteristics of an effective language teacher from not only FL teachers but also FL students. Two hundred teachers and 406 ninth-grade students in Tel Aviv, Israel were "randomly selected" (p. 129) to complete the survey. Unlike Reber's (2001) instrument, which used Likert-type questions, Brosh (1996) provided teachers and students with the same list of 20 characteristics and asked them to choose the three most important and to rank order those three. The first three items are included to give the reader a sense of the instrument: 1.) Prepares and organizes the lesson, 2.) Acquainted with the curriculum, 3.) Helps students after class time (p. 136). Of the 20 total items, only five were specific to language learning.

The results of this specific questionnaire demonstrated that students' and teachers' perceptions were largely homogeneous. The item that both groups chose as most important was the teacher's command of the target language. The second most crucial factor for both students and teachers was the instructor's ability to transmit knowledge comprehensibly while motivating students to do their best. Students' choice for the third most important characteristic differed with teachers' as they indicated the essential need of being treated equitably and fairly while teachers ranked the ability to provide students with successful experiences as third.

These studies provide valuable information in regard to what language teachers and students think, in general, about what effective teaching is, but both lacked a crucial component. In both cases, the students and teachers were not reacting to an actual demonstration of teaching and providing an evaluation of that teaching. Moore (1996) takes a step in that direction by administering the exact same evaluation instrument to both teachers and their students.

As part of his dissertation research, Moore (1996) investigated the correlation between teachers' and students' perceptions of teaching effectiveness. He specifies the purpose of his research by stating that it was "to evaluate graduate teaching assistants' perceptions of their teaching effectiveness and correlate these perceptions with students' evaluations of graduate teaching assistants' performance" (p. 4). Moore's study included 129 graduate teaching assistants and their 3,088 students from various departments across a large university in the southeastern United States. The results indicated that students were consistent in evaluating their graduate student instructors lower than the instructors rated themselves on the exact same instrument with the wording slightly altered to fit each of the two groups. A positive correlation was also found between prior teaching preparation and experience and the perceptions of students and teachers. Although this study was not conducted exclusively in the language classroom, it demonstrates the disconnect that may occur between teachers and students regarding an assessment of teaching.

Differing perceptions manifest themselves not only on attitudinal surveys, but also on teacher-course evaluations. In her historical survey of FL teacher development, Schulz (2000) reports on the conclusions of a task force for the University of California system which found student evaluations to be the "predominant method of evaluating teaching" (Schulz, 2000, p. 511). Likewise, Pennington and Young (1989) argue that student ratings many times "provide input for a summative evaluation process related to employment actions such as contract renewals, tenure, or promotion" (p. 627). Many teachers are naturally aware of this and as Bernhardt (2001) observes, "teachers often equate 'effective' with positive student evaluations. This equation is not necessarily meaningful or appropriate, but it is pervasive" (p. 47). The assumption underlying the use of student evaluations is that students are capable of making valid, useful assessments of their teachers' instruction. With the increased integration of student evaluations in professional advancement and the weighty consequences for teachers, the question remains, "How valid and reliable are student evaluations of teaching?"

One of the foremost authorities on student evaluation research, Aleamoni (1981), identifies the rationale behind the use of student evaluations. He outlines four main justifications: 1) students represent the main source of evidence regarding the accomplishment of educational goals, i.e. motivation, rapport between teacher and student, 2) in contrast to outside observers, students interact directly with the text, the course content, the method of instruction, all of which affect student attitudes and achievement, 3) student evaluations facilitate communication between teacher and student, especially in large classes, and 4) student evaluations constitute a means by which other students may base their selection of courses and instructors.

Aleamoni continues his discussion of student evaluations by articulating common concerns of teachers, listed here: 1-inconsistency of student evaluations due to immaturity and lack of experience; 2-most student ratings of professors reflect a popularity contest; 3-student rating forms are both unreliable and invalid; 4-other variables often cited: size of the class, sex of the student and instructor, time of day the course was offered, etc., and 5-students' grades are highly correlated with their ratings of the course and the instructor. In a later publication, Aleamoni (1987) addresses all these concerns by reviewing relevant research. Aleamoni systematically refutes each concern citing research that debunks these eight fears expressed by teachers. The author emphasizes how crucial it is that professionally constructed instruments be used in collecting student ratings.

Contrary to Aleamoni's (1987) conclusions, a study by Wennerstrom and Heiser (1992) presents evidence that ESL student evaluations are systematically biased. They found significant effects for ethnic background, level of English, course content, and attitude toward the class on an instrument they administered to 522 ESL university students in an academic program and to 2,658 students in an intensive English program. Indonesians, Chinese, Latin Americans, and Arabic students rated higher on average than the Japanese students. In addition, higher-level students rated lower than did lower level students. Finally, older students rated slightly higher than younger ones. In discussing their findings, the authors wisely note the particulars of their program that may have influenced the results. For example, the intensive program had a conference-centered writing track which may have had a negative impact on the evaluations of writing courses. In addition, the course content varied according to level for both programs and that may have had an impact. The impact of local context and curriculum may make a significant difference in the evaluation of teaching.

In a lengthy and detailed analysis of the reliability and validity of student ratings, Feldman (1998) astutely concludes "we do not, in fact, know very much about what does go on in students' minds when they fill out rating forms" (p. 51). Although it may not be entirely clear what exactly students take into consideration in filling out teacher evaluations, there appears to be evidence that students do reflect on the use of their evaluations by faculty and administration. Spencer and Schmelkin (2002) found that students were generally unconcerned about negative repercussions for filling out evaluations.

They discovered that students who felt positively about teaching in general also felt optimistic that their evaluations would be considered.

A critical issue is whether students' ratings are valid since discussions of student ratings and evaluations generally revolve around issues of promotion, rank advancement, i.e. summative evaluation vs. formative evaluation. Berman (2003) identifies several characteristics of effective formative and summative evaluation. She specifies the need for evaluations to be multi-faceted, supportive of collegiality, and faculty-driven.

Students do not represent the sole source of teacher evaluations and impetus for professional development; other viable sources of evaluation may be teacher interviews, student achievement, classroom observation, peer review, and faculty self-evaluation (Pennington & Young, 1989). As Berman (2003) perceptively observes, evaluation systems need to be faculty-driven and that teacher must feel supported by other colleagues. Recently, Bailey, Curtis, and Nunan (2001) have published an entire textbook devoted to the professional development of language teachers through reflective means. Bailey et al. recommend that teachers look no farther than themselves as a starting point in their never-ending quest for 'effective teaching.' However, reflective teaching should not be considered an individual endeavor bereft of collegial support, encouragement, and direction. Accordingly, Bailey et al. include chapters on peer observation, team teaching, and mentoring and coaching.

Pennington and Young (1989) sort through the research on different techniques of self-evaluation and reflective teaching to conclude that the main benefits of self-evaluation for teachers are the strong probability of spurring change, the potential for encouraging a sense of responsibility and professionalism, and the opportunity to focus on long-term goals for the individual teacher and the overall program of which the teacher forms a part. However, the authors note that self-evaluation usually lacks reliability and objectivity for summative evaluation, and even for formative evaluation this reflective approach may not be valid as "insecure teachers tend to overrate themselves, and secure teachers tend to underrate themselves" (p. 640). The feedback from a peer or a supervisor may prove invaluable in ascertaining a teacher's effectiveness and overall ability.

The final approach to determining teacher effectiveness relevant to the study at hand concerns peer evaluation through observation. Similar to the previous two approaches to teacher evaluation, this approach has its merits and its faults. Bailey et al. (2001) laud the advantages of peer observation and define it as "the act of being openly and attentively present in another's classroom, watching and listening to the classroom interaction primarily for reasons of professional growth (rather than supervision or evaluation)" (p. 157). They state that peer observation is not the traditional expert-novice relationship present in many professional development programs. Bailey et al. claim that peer observers may benefit as much, or more, than the observed. Nonetheless, others have observed that the inclusion specifically of a peer versus a detached supervisor in evaluative practices brings its own difficulties. Some of these include the fear of harming working relationships, the lack of sufficient time, and uncontrollable variance among different peers in spite of the instrument used. (Pennington & Young, 1989).

The concerns and challenges of self-evaluation and peer observation, as outlined by Pennington and Young (1989), were given in the context of faculty *evaluation* not professional *development*. Issues of evaluation inevitably have professional consequences, i.e. promotion, pay increase, and, hence, are more sensitive than issues exclusively pertaining to development, although the two may be closely intertwined. For the purposes of the current study which addresses self-evaluation, peer observation/evaluation, and student evaluation, the concerns merit mention. The authors recommend that peers should observe several times over a period of time and should be adequately trained with experience teaching the classes they observe.

In summary, this review of literature has offered an overview of research, standards, and teaching evaluation instruments that have helped shape current thinking on what is effective FL and SL teaching. Second, the review has addressed the issue of student perceptions and evaluations of teaching. Finally, some pros and cons of self-evaluation and peer observation have been presented. This background will better inform the reader as to the relevant issues surrounding the current research outlined below.

STUDY

Participants

The following study was conducted in a major university in the southwestern United States. One Graduate Associate in Teaching (GAT) in the Spanish and Portuguese Department, a Ph.D. student at the university, took part in this study, as did the students from the GAT's two classes. The criterion for selection of the instructor was that the person had taught the same course before so as to control for the effect a new curriculum and syllabus would have on the instructor's teaching effectiveness. Furthermore, this instructor was selected based on how the researchers' schedules coincided with the two classes taught. The instructor was a native-speaker of Spanish completing graduate studies in applied linguistics at the same university where the classes were taught. The GAT had three semesters of experience teaching Spanish as a FL and ten semesters teaching English as a Foreign Language (EFL) in Latin America.

Of the 47 students registered in the two classes, a total of 39 completed the two phases of the study. Forty-three participated in the first study and 42 in the second, of which 21 were female and 22 were male. All of them were 25 years of age and under. However, nine of the participants did not provide their age in the demographic information. All participants that took part in this study did so voluntarily.

The researchers who observed and evaluated the teaching of the instructor are doctoral students in Second Language Acquisition and Teaching majoring in pedagogy. Two of them have master's degrees in Spanish Pedagogy and one of them in teaching English as a Second Language (ESL). All of them had received instruction in teacher training and were experienced teachers of Spanish (4-6 years). All of the researchers were also fluent speakers of Spanish.

Methods

This study employed evaluative questionnaires as the main method of data collection. In order to make valid evaluations of the teacher's effectiveness, the three researchers attended both of the selected classes for a total of seven observations over the span of three weeks. During each visit, the observers took extensive field notes with the purpose of documenting all aspects relevant to the completion of the teacher-effectiveness questionnaire.

Data for the questionnaires were collected in two phases. The first phase took place during the first week of the study. Participants were asked to complete Questionnaire #1 (see Appendix A) after the first observation. The second phase took place in the third week of the study during the last observation. The instructor, the students, and the observers completed Questionnaire #2 (see Appendix B). The teachers and students filled out the questionnaire based on their participation in the language classroom while the researchers completed the same questionnaire based on their observations. During the entire observation process, the researchers did not discuss or share comments on the lessons observed with each other, the instructor or the students, nor did they have access to any of the students' evaluations until their evaluations were complete. In addition, the teacher and the students completed a demographic questionnaire.

Data collection instruments

Demographic questionnaires: Participants completed a background questionnaire in the last phase of the study. Besides providing demographic information, participants answered items about their experiences with foreign languages. The purpose of the questionnaire was to find out the number of years of previous language exposure as well as the overall quality of their prior FL learning experience.

Teacher-Course Evaluation (TCE) questionnaire (see Appendix A): Participants were surveyed on their instructor's teaching effectiveness. Five items out of the seventeen on the official TCE forms were chosen. These forms are used in many departments throughout the university, including Spanish and Portuguese, as one data source to evaluate instructors' teaching effectiveness. The items chosen targeted directly the students' perceptions of their teacher's effectiveness. The items excluded were more related to departmental concerns such as the quality of textbooks, the amount of homework given, the overall difficulty of the class, etc. After responding to the Likert-scale items, participants were asked to explain in an open-ended format what factors they had taken into account rating the instructor's teaching effectiveness.

Teacher-effectiveness questionnaires (see Appendix B): The questionnaire was developed from the compilation of three currently available instruments for evaluating teachers' effectiveness. One by Reber (2001) was based on the literature on FL teacher effectiveness. The second one was developed by Moore (1996) to compare teachers' and students' perspectives on teacher effectiveness. The third instrument consulted by the researchers was the *Characteristics of Effective FL Instruction* developed by the National Association of District Supervisors of Foreign Languages (1999). The first step in the preparation of the questionnaire was to determine the content that needed to be included. Nine categories (Table 1) pertaining to teacher effectiveness were included. The second step was to select and adapt questions from the consulted instruments in order to cover the nine categories.

Table 1: Breakdown of categories and questionnaire items from survey

Categories	Questionnaire items
1. Command/use of target language	Items 1, 2, 3 and 35.
2. Student involvement	Items 5, 7 and 8.
3.Variety/appropriateness of learning activities	Items 9, 10, 11 and 12.
4. Clarity and brevity of instructions and grammar explanations	Items 13, 14, 15 and 17.
5. Appropriate use of materials and realia	Items 18, 19 and 20.
6. Communicative interaction in the target language	Items 6, 7, 21, 22, 23 and 24.
7. Efficient use of time	Items 4, 24, 25, 26, 27 and 28.
8. Rapport with students/personality factors	Items 16, 29, 30, 31 and 32.
9. Culture	Items 33, 34 and 35.

The same instrument was used for the students and the peer observers while the teacher's instrument was only slightly altered to read "I . . ." versus "The teacher . . . ." The instrument included thirty-five items administered by way of a four-point Likert-type scale ranging from strongly agree (4) to strongly disagree (1).

RESULTS

The results showed that there were significant differences between the students, the teacher, and the third party observers in many of the areas analyzed. In order to answer the research questions, two tailed *t*-tests were used to determine whether resulting differences between the groups were significant. Table 2 shows the mean scores on Questionnaire #2 of each evaluation groups' overall rating of the teacher's effectiveness as well as the teacher's self-evaluation. Class #1 gave the highest overall score on the effectiveness of the teaching, 3.54 while the teacher gave the lowest evaluation of the effectiveness of the teaching, 2.83.

Table 2: Means of Questionnaire #2 by the Different Evaluation Groups

Evaluation Groups	Mean Questionnaire #2 (max. 4.0 and min. 1.0)
Class #1	3.54
Class #2	3.35
Peer Observers	3.26
Teacher	2.83

In answer to the first research question regarding student's perceptions of effective teaching versus those of their teacher, Table 3 shows the results of the students' evaluation of their teacher and the teacher's self-evaluation. In comparing the teacher with all three groups, a significant difference was found ($p < .05$) between the teacher's view of her teaching effectiveness and how the students viewed the effectiveness of her teaching.

Table 3: Students' Evaluations vs. Teacher's Self-Evaluation on Questionnaire #2

Evaluation Groups	*t*-Test Results of Significance ($p < .05$)
Class #1 vs. Teacher's Self-Evaluation	$p < .001$
Class #2 vs. Teacher's Self-Evaluation	$p < .001$
All Students (Class #1 & #2) vs. Teacher's Self-Evaluation	$p < .001$

When comparing the results of the evaluation groups as stated in research question #2, students' and teacher's evaluations of effective teaching were analyzed against those of the peer observers. Using *t*-tests to test significance, the results showed that there were significant differences between the peer observers, Class #1, the teacher, and all of the students (Class #1 & #2). However, Class #2 did not show a significant difference when compared to the results of the peer observers. These results can be seen in Table 4.

Table 4: Peer Evaluations vs. Other Groups' Evaluations of Teacher Effectiveness on Questionnaire #2

Evaluation Groups	*t*-Test Results of Significance ($p < .05$)
Class #1 vs. Peer Observers	$P < .001$
Class #2 vs. Peer Observers	$P = .18$
Teacher vs. Peer Observers	$P = .002$
All Students (Class #1 & #2) vs. Peer Observers	$p = .005$

To answer research question #3, the researchers compared the results of the five questions on the TCE questionnaire to the results of Questionnaire #2 to test whether there was a significant difference between the two instruments, both of which purport to measure teacher effectiveness. The TCE consisted of a 5-point Likert-type scale that was converted to a four-point scale for the purpose of comparing the data using two-sample *t*-tests. The results are found in Table 5, which shows a significant difference between Class #1 and Class #2 when compared with the questions from the TCE with the students from both classes rating the teacher higher (3.54 and 3.35) on Questionnaire #2 than they did on the TCE (3.19 and 3.06). Class #1 rated the teacher higher on both instruments.

Table 5: Comparison of the TCE with Questionnaire #2 on Teacher Effectiveness

Instruments	T-Test Results of Significance($p \leq .05$)
Class #1 Results of Questionnaire #2 vs. TCE	p= .002
Class #2 Results of Questionnaire #2 vs. TCE	p= .01
Class #1 and #2 Results of Questionnaire #2 vs. TCE	p= .003

In answering research question #4 regarding what the students took into account when they filled out the university's TCE, the qualitative portion of the instrument was included to ascertain why students rated their teacher as they did on the five questions from the TCE. A total of 215 comments were written describing the reason for the evaluation, some of these are included below. Of the 215 comments, 185 (86%) were determined by the researchers to be positive comments whereas 30 (14%) were classified as negative comments in determining the evaluation rating.

The first question asked to the students from the TCE was "What is your overall rating of this instructor's effectiveness?" More than 90% of the students rated the teacher a 4 (usually effective) or a 5 (almost always effective) on a scale of 5 (see Appendix A). All of the students rated their teacher a 3 (sometimes effective) or above on this question. The most common reason the students gave (39%) dealt with the clarity of the teachers' explanations and instructions. Some typical examples from the students were "Thorough and good explanations" and "Makes Spanish understandable." Additionally, many of the students (17%) cited personality traits as an important factor. Other categories that students included in their positive rating of their teacher were type of activities, amount learned, and the teacher being demanding. Those who put a negative rating cited such issues as use of too much Spanish and not enough explanations.

The second question from the TCE dealing with effective teaching was "How much do you feel you have learned in this course?" More than 70%

of the students rated the teacher as a 4 (more than usual) or a 5 (an exceptional amount) on a 5-point scale. The most common explanations for the score (51%) that the students gave dealt with a comparison of previous experience and knowledge with what they felt they had learned in the current class. Some typical examples are "More than in all my high school classes" and "Reviewed and really grasped previously learned concepts." Many of the students made positive and negative comments relative to the same issues i.e. some students felt review was helpful while others did not.

The third question given to the students was "What is your overall rating of this course?" Whereas the other responses were limited in their scope, this question offered the widest variety of responses in part because of the broad nature of the question. The notion of the class being fun or entertaining was a reason given by 22% of the students and 20% stated that the amount learned contributed to their positive rating. Personality, activities, and effective teaching were also given as reasons for the positive score. Some examples are "I have fun while learning," and "Learned all skills reading, writing, speaking and listening." Student complaints about the course and subsequent low rankings of their teacher stems from a variety of factors such as "Boring, repetitive," "Pace too fast," etc. One comment made by several students relates to the Spanish Department's grading policy which sets an A at 92% instead of 90% like other departments.

The fourth question asked was "Rate the usefulness of the in-class activities (lectures, discussions, etc.) in this course in helping you learn?" This question directly deals with effective teaching by asking what goes on in the classroom. The most common reason (37%) given for a positive rating (4 or 5) was the usefulness of the material and 73% of the students gave the teacher a 4 (usually effective) or a 5 (almost always useful). Several students simply responded "Useful" where as others stated "Understand concepts better." The second most common response (24%) referred to the type of activities used in the class such as "Games are helpful," "She makes us talk," and "Hands on activities." The negative comments given by the students were almost the exact opposite of the positive ones that were received: "Boring activities," "Repetitive activities," and "Sometimes we don't understand what we are supposed to be doing".

The fifth and final question to which the students were asked to respond was "What is your rating of this instructor compared with other instructors you have had?" Over 95% of the students rated their teacher from 3 (about as effective as most) to 5 (one of the most effective). Also, almost half of the students (44%) gave their instructor a 5, which is a higher proportion than in any of the other five TCE questions. As would be expected from the question, most students based their comments on comparisons between their current teacher and previous ones. Some of the representative comments from the students were "She is my favorite/best Spanish teacher," "My favorite college teacher," "The only one I actually learned from," etc. Personality factors were again mentioned as a major component in the ranking of the instructor. Remarks such as "Nice person" and "She cares and wants you to learn" were all common explanations for the justification of the score. Additionally, teacher effectiveness was cited as a reason for the positive evaluation. No negative comments on this question were given regarding the instructor.

In an attempt to answer the research questions, statistical tests of significance and qualitative data were accompanied by several other tests to explain the differences between the groups. When gender was investigated, the researchers found a significant difference (p= .01) between the male (n=22) and the female (n=21) students in their evaluation of the effectiveness of their teacher. The mean score of the male students was lower than that of the female students. Two-sample *t*-tests were run on each individual question to see if significant gender differences could be found in answering the question from Questionnaire #2 (see Appendix B). Significant differences (p<.05) were found on two questions when analyzed according to gender. They were "makes use of activities that are appropriate for learning the language" and "monitors students' progress during activities." The male students ranked the teacher lower on both questions.

The researchers also ran tests to determine if there were significant differences in the individual responses from Questionnaire #2 of Class #1 and Class #2. When comparing the results of the two classes, a significant difference between them was found in 8 (23%) of the 35 questions. Three of the areas where significant differences were found were communicative interaction in the target language, efficient use of time, and implementation of culture in the classroom. Class #2 consistently rated the instructor lower than Class #1 in these three areas.

DISCUSSION

Several conclusions can be drawn from the data that were gathered and analyzed. Of all the statistical analyses run between multiple combinations of groups and instruments, the most telling finding is the fact that only one combination did not show a significant difference on one instrument, i.e. the peer observers and Class #2 on Questionnaire #2. This supports previous research (Moore, 1996) that found that there were significant differences between teachers' views of effective teaching and their own students' view of those same teachers' effectiveness. The addition of peer evaluators produced a third and significantly different view on the effectiveness of the teacher when compared to the teacher and Class #1.

During the observation period, the peer evaluators did not notice any significant difference in the teaching and instruction from one class to another. However, the personality and make up of the students in each class is reflected in their evaluations of the instructor. It is also important to note that Class #2 was the second class that was taught by the instructor and it was taught immediately after Class #1. This might lead one to believe that the teacher learned from the mistakes made in the first class and improved on them in the second class; nevertheless, the evaluations were still lower in the second class.

Anecdotally, the instructor mentioned to the researchers after the completion of data collection that Class #1 was more difficult than the second one and that Class #2 was more advanced than Class #1. The instructor's observation that the two classes represented different personalities was corroborated by the study's results lending credence to the need to evaluate a teacher based on more than one class's evaluations.

Of the four groups analyzed—peer observers, Class #1, Class #2, instructor—the instructor gave the teaching the lowest rating. This was not entirely unexpected given the education of the instructor. This teacher is enrolled in a doctoral program that focuses on language pedagogy and thus is aware of the high standards that are set in language teaching, such as culture, target language use, group activities, etc. and how these make up an effective SL classroom. This awareness may have contributed to the low self-evaluation of the teaching effectiveness. Furthermore, some teachers are more self-critical than others which may have been the case in the current study. The groups that gave the highest ratings were the two classes, which is contrary to the results of Moore (1996) who found that students evaluations of their teachers were lower than the teachers' evaluations. This raises again the question of using the students' evaluation as an instrument upon which high stakes decisions are based. The idea that student evaluations are not accurate is not a novel concept but in which direction the discrepancy occurs may not only be in the negative; students may overrate their teachers.

The peer evaluators showed some difference in their responses to the questionnaire. This offers insight into the notion that educated specialists in the field of pedagogy and teaching evaluation also show differences in opinion. One way that the peer evaluators tried to mitigate these minimal differences was by arriving at a consensus on each item. After a consensus was reached, the overall mean of the consensus and the individual results were taken. The researchers found that a difference of only .02 existed between the consensus mean (3.23) and the mean of each of the individual scores (3.25); hence, the individual results were used in the data analysis. Regarding the overall agreement on the individual questions on Questionnaire #2, a 49% correlation existed between the peer evaluators on the thirty-five questions. All of the questions except one had agreement between at least two of the researchers. Only three of the questions had differences of more than one point on the 4-point Likert-type scale. While these results reflect the differences in perspective that peer evaluators have, overall they were not significant.

The TCE is used university wide in all departments to evaluate instructors and the results affect teachers both summatively and formatively. The researchers found a significant difference between the scores on the 5-questions dealing with teacher effectiveness and Questionnaire #2, which was designed as a more specific instrument to measure the effectiveness of language teachers versus instructors of other disciplines. When the students were asked in Questionnaire #2 to evaluate the effectiveness of their instructor, they rated their teacher higher in both classes than on the five broad questions from the TCE. This provides evidence that a discrepancy may exist when teachers are evaluated with a more specific instrument designed for language teaching versus a general instrument that is used throughout an entire university.

As the students explained their rationale behind the scores given to their teacher regarding the 5-questions from the TCE, the students considered a wide variety of factors in their evaluation of their instructor and the course. Various students mentioned factors that were not related to the teacher or the classes. Issues such as previous experience, grading policies, university requirements all influenced the students scoring of the teacher's effectiveness and the overall evaluation of the class. Additionally, students' scores on the evaluation did not always logically concur with their explanation, that is to say that several times negative or neutral comments were included and yet a high score (4 or 5) was given. Conversely, many of the comments were positive and yet resulted in a low or average score (1, 2, or 3). Some of these factors, as mentioned previously, dealt with the issue of teaching and teacher effectiveness. However, many of them did not and yet adversely influenced the student's response. This study offers more evidence toward the abstract idea of an ideal "teacher personality." This was one of the factors considered by some students on every question from the TCE.

Finally, the students were asked to provide an overall evaluation of the quality of their previous experience with language teaching. The researchers hypothesized that students' previous language learning experiences may have had a significant effect on their current perceptions and evaluations. A *t*-test was run to investigate whether the students' responses to this question differed from their evaluation of their current class in question #3 of the five TCE questions (see Appendix A). No significant difference between the overall rating of their current course and their previous experience with FL study was found. The mean of both the previous experience (2.86) and the students' current course rating (2.96) varied by .1 on a four-point scale.

Further analysis demonstrated that students' assessments of their teacher's effectiveness and the overall quality of the course were significantly different. Students evaluated their teacher's effectiveness at 3.44 on a 4-point scale as measured in Questionnaire #2 and lower, 3.13, than the overall score given on the five questions from the TCE. As mentioned above the overall course rating was much lower 2.96 on a 4-point scale. Evidently, students may rate the course and teacher at different levels and if these ratings are averaged together the resulting mean may not necessarily give an accurate rating of the teacher's performance. This discrepancy is particularly important to any teacher who does not have the power to select their curriculum or syllabus. As seen through the students' comments, factors beyond the instructor's control were taken into account in the evaluations and the instructor was penalized for these issues.

CONCLUSIONS

This research not only brings into question the validity of student evaluations but also raises concerns regarding both peer and self-evaluations. The data support using multiple perspectives in the evaluation of any teacher. Through these multiple perspectives, an approximation of a teacher's effectiveness can be attained. This study also presents two classes taught by the same instructor at the same time of day and yet with significantly different results. Future research needs to provide additional studies that investigate why students evaluate teachers the way they do and what factors are taken into account during these evaluations. This information would help teachers provide students with a positive experience and also would help in program development. A teacher who was informed with this information might be able to mitigate some of the problems and difficulties that students have in beginning language classes.

The researchers also showed that when the instrument is designed and developed specifically for language teaching as compared to general evaluative instruments, significantly different results occur. Each department or college should develop instruments that more finely measure the characteristics of effective teaching in subject-specific environments. By using an instrument that reflects the characteristics of a given field of teaching, more specific problems could be addressed rather than a global response dealing with overall class or instructor rating. A specific instrument also can help to focus the students on the different aspects of the teacher's effectiveness and contribute to greater accuracy in their evaluation.

Given that beginning Spanish is a general education (GE) requirement, the classes consisted of many students who were obligated to take the course to graduate. A negative attitude toward the class and language possibly could be attributed to this factor. If an instructor receives low evaluations in teaching GE courses, the explanation may not be any more complex than the simple fact that many students in GE courses may have very low motivation.

Regarding self-assessment, the researchers found that this instructor gave the lowest evaluation of all groups that participated. The affect of training and the knowledge of sound pedagogical principles may lead very well lead to these types of lower self-evaluations where as the uninformed teacher may be self-aggrandizing due to a lack of knowledge. In conclusion, the aforementioned concept of using multiple perspectives in evaluation is reinforced in this study and future research into the best way to use multiple perspectives is needed.

REFERENCES

Aleamoni, L.M. (1981). Student ratings and instruction. In J. Millman (Ed.), *Handbook of Teacher Evaluation* (pp. 110-145). Beverly Hills: Sage.

Aleamoni, L.M. (1987). Typical faculty concerns about student evaluation of teaching. In L.M. Aleamoni (Ed.), *Techniques for Evaluating and Improving Instruction* (pp. 25-31). San Francisco: Jossey-Bass.

Bailey, K.M., Curtis, A. & Nunan, D. (2001). *Pursuing Professional Development*. Boston: Heinle & Heinle.

Berman, E. Characteristics of Good Evaluation Systems. Class Handout in SLAT 596B, University of Arizona, Fall, 2003.

Bernhardt, E.B. (2001). The professional development of highly experienced and less experienced teachers: Meeting diverse needs. In B. Rifkin (Ed.), *Mentoring Foreign Language Teaching Assistants, Lecturers, and Adjunct Faculty* (pp. 41-53). Boston: Heinle & Heinle.

Brosh, H. (1996). Perceived characteristics of the effective language teacher. *Foreign Language Annals, 29*, 2, 125-138.

Feldman, K.A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In Smart, J. (Ed.), *Higher Education: Handbook of Theory and Research* (pp. 35-74). New York: Agathon Press.

Horwitz, E.K. (1990). Attending to the affective domain in the foreign language classroom. In S. Magnan (Ed.), *Shifting the Instructional Focus to the Learner* (15-33). Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages.

Mitchell, C.B., &Vidal, K.E. (2001). Weighing the ways of the flow: Twentieth century language instruction. *The Modern Language Journal, 85*, 1, 26-38.

Moore, M.S. (1996). *Assessing college teaching effectiveness: A comparison of graduate teaching assistants and their students.* Unpublished Doctoral Dissertation. DAI, 57, no. 05A, University of Auburn.

Pennington, M.C., & Young, A.L. (1989). Approaches to faculty evaluation for ESL. *TESOL Quarterly, 23*, 4, 619-646.

Reber, T. (2001). *Effective teaching behaviors and attitudes as perceived by foreign language teachers.* Unpublished Doctoral Dissertation. University of Arizona.

Schulz, R.A. (1996). Focus on form in the FL classroom: Students' and teachers' views on error correction and the role of grammar. *Foreign Language Annals, 29*, 3, 343-364.

Schulz, R.A. (2000). Foreign language teacher development: MLJ perspectives—1916-1999. *The Modern Language Journal, 84*, 4, 495-522.

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

Stevens, J. (1987). Using student ratings to improve instruction. In L. Aleamoni (Ed.), *Techniques for Evaluating and Improving Instruction* (pp. 33-46). San Francisco: Jossey-Bass.

University of Arkansas, Department of Foreign Languages Assessment Instrument. Handout from SLAT 596B, Fall Semester, 2003.

Virginia Beach City Public Schools (2000). Evaluation Form for Foreign Language Teachers. Handout from SLAT 596B, Fall Semester, 2003.

Wennerstrom, A.K., & Heiser, P. (1992). ESL student bias in instructional evaluation. *TESOL Quarterly, 26*, 2, 271-288.

APPENDIX A

<u>Questionnaire #1: Five Questions from TCE</u>
<u>Teacher-Course Evaluations</u>

Place a circle around the phrase that best describes your opinion regarding the following five questions.

1. What is your overall rating of this instructor's teaching effectiveness?

almost always effective -	*usually effective -*	*sometimes effective -*	*rarely effective -*	*never effective*
5	4	3	2	1

2. How much do you feel you have learned in this course?

an exceptional amount-	*more than usual-*	*about as much as usual-*	*less than usual-*	*almost nothing*
5	4	3	2	1

3. What is your overall rating of this course?

one of the best-	*better than average-*	*about average-*	*worse than average-*	*one of the worst*
5	4	3	2	1

4. Rate the usefulness of the in-class activities (lectures, discussions, etc.) in this course in helping you learn?

almost always useful -	*usually useful -*	*sometimes useful -*	*rarely useful -*	*almost never useful*
5	4	3	2	1

5. What is your rating of this instructor compared with other instructors you have had?

one of the most effective	*more effective than most*	*about as effective as most*	*less effective than most*	*one of the least effective*
5	4	3	2	1

What did you take into account when you answered the following questions on the previous page. Please note down your thoughts on the lines provided under each question. Try to be as thorough as possible in reflecting on why you chose the ratings you did.

1.What is your overall rating of this instructor's teaching effectiveness?
# Rating: ________

2. How much do you feel you have learned in this course?
# Rating:________

3. What is your overall rating of this course?
# Rating:________

4. Rate the usefulness of the in-class activities (lectures, discussions, etc.) in this course in helping you learn?
# Rating:________

5. What is your rating of this instructor compared with other instructors you have had?
#Rating:________

APPENDIX B

Questionnaire #2: Instructor/Students' Questionnaire
The Effective Foreign Language Teacher

We are conducting a study to investigate how multiple perspectives on effective teaching compare. We would like you to help us by completing this survey concerning your experiences in your Spanish classroom. There are no right or wrong answers. Right answers are the ones that are true for you. Please respond to answers sincerely as only this will guarantee the success of the investigation. **Thank you!**

Instructions: Please carefully read each statement and indicate to what extent you agree or disagree by

Circling the number that best describes your opinion.
4-Strongly Agree 3-Agree 2-Disagree 1-Strongly Disagree

I/The instructor . . .				
1-use/s language that is comprehensible to students.	4	3	2	1
2-use/s the foreign language as the predominant means of communication in the classroom.	4	3	2	1
3-use/s the foreign language competently.	4	3	2	1
4-show/s evidence of planning in each class.	4	3	2	1
5-use/s small groups and pair work to help learners experience a greater degree of involvement.	4	3	2	1
6-provide/s sufficient opportunities for students to practice Spanish.	4	3	2	1
7-use/s effective questioning techniques to elicit responses from students.	4	3	2	1
8-encourage/s all students to participate in the classroom.	4	3	2	1
9-provide/s opportunities for extensive practice of and/or exposure to each grammatical structure or topic being presented.	4	3	2	1
10-integrate/s a variety of activities that appeal to different learning styles (i.e., kinesthetic, visual, auditory, tactile) as well as different interest areas (i.e., sports, music, etc.).	4	3	2	1
11-make/s use of activities that are appropriate for learning the language.	4	3	2	1
12-sequence/s activities from easy to difficult.	4	3	2	1
13-teach/es grammar in a clear, concise, and understandable manner.	4	3	2	1

I/The instructor . . .				
14-assure/s that students understand activities fully	4	3	2	1
before starting.	4	3	2	1
15-answers questions precisely.				
16-demonstrate/s respect for students.	4	3	2	1

4-Strongly Agree 3-Agree 2-Disagree 1-Strongly Disagree

I/The instructor . . .				
17-provide/s sufficient number of examples to illustrate	4	3	2	1
explanations.				
18-use/s the textbook wisely without boring students.	4	3	2	1
19-frequently uses supplemental materials and visual	4	3	2	1
aids to enhance instruction.				
20-use/s the blackboard, VCR/TV, and overhead	4	3	2	1
appropriately in order to aid instructional goals.				
21-create/s a comfortable and accepting atmosphere for	4	3	2	1
students to speak the foreign language.				
22-allow/s students ample opportunity to engage in	4	3	2	1
meaningful and communicative interactions in the				
foreign language.	4	3	2	1
23-correct/s students effectively in a non-threatening,	4	3	2	1
encouraging manner.				
24-find/s an appropriate balance between the time the	4	3	2	1
teacher spends talking and the time allowed for students	4	3	2	1
to talk.	4	3	2	1
25-maintain/s an adequate pace of activities.				
26-transition/s smoothly from one activity to another.	4	3	2	1
I/The instructor . . .	4	3	2	1
27-make/s efficient use of class time by allocating	4	3	2	1
appropriate time for each activity.	4	3	2	1
28- monitor/s students' progress during activities.	4	3	2	1
I/The instructor . . .				

29-am/is interested in and accessible to individual students' needs.	4	3	2	1
30-demonstrate//s enthusiasm and a notable energy in teaching the foreign language which motivates students to learn.	4	3	2	1
31-demonstrate/s a friendly and patient disposition.	4	3	2	1
32-praise/s students effectively to acknowledge students' achievements.	4	3	2	1
33-integrate/s culture frequently in daily classroom activities.	4	3	2	1
34-demonstrate/s a positive attitude toward cultural diversity.	4	3	2	1
35-have/has an extensive knowledge of the Spanish language and culture.	4	3	2	1