

To my beloved husband

DESIGNING A VALID EFL-SPECIFIC TEACHER EVALUATION FORM
FOR STUDENTS' RATINGS

The Graduate School of Education
of
Bilkent University

by

AYÇA ÖZÇINAR

In Partial Fulfillment of the Requirements for the Degree of
Masters of Arts

in

The Department of
Teaching English as a Foreign Language
Bilkent University
Ankara

July 2011

BİLKENT UNIVERSITY
THE GRADUATE SCHOOL OF EDUCATION
MA THESIS EXAMINATION RESULT FORM

July, 2011

The examining committee appointed by the Graduate School of Education
for the thesis examination of the MA TEFL student

Ayça Özçınar

has read the thesis of the student.

The committee has decided that the thesis of the student is satisfactory.

Thesis Title: Designing a valid EFL-specific teacher
evaluation form for students' ratings

Thesis Advisor: Vis. Asst. Prof. Dr. JoDee Walters
Bilkent University, MA TEFL Program

Committee Members: Asst. Prof. Dr. Julie Mathews Aydınlı
Bilkent University, MA TEFL Program

Lecturer Seniye Vural
Erciyes University

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

(Vis.Asst. Prof. Dr. JoDee Walters)
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

(Asst. Prof. Dr. Julie Mathews Aydımlı)
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

(Lecturer Seniye Vural)
Examining Committee Member

Approval of the Graduate School of Education

(Vis. Prof. Dr. Margaret Sands)
Director

ABSTRACT

DESIGNING A VALID EFL-SPECIFIC TEACHER EVALUATION FORM
FOR STUDENTS' RATINGS

Ayça Özçınar

M.A. Program of Teaching English as a Foreign Language

Supervisor: Vis. Asst. Prof. Dr. JoDee Walters

July 2011

Teacher evaluation forms (TEF) for students' ratings have been used since the 1920s. Many researchers have been questioning the validity of such forms since then. There have been many attempts to design field-specific TEFs in several studies.

The present study aims to design a valid EFL-specific TEF for students' ratings. The instruments of the study are an online questionnaire, three versions of the teacher evaluation form and the form itself (the final version of the form). The participants are 392 students from different proficiency levels (elementary, lower intermediate, intermediate, upper intermediate and advanced) and 21 teachers at Anadolu University School of Foreign Languages (AUSFL), and three teachers from different universities. The information gathered from the online questionnaire and the semi-structured interviews on the three versions of the form was analyzed qualitatively, and a final version of the form was designed. The piloting of the final version of the form was analyzed quantitatively.

The qualitative analysis showed that the current TEF for students' ratings at AUSFL was not satisfactory, and the literature on effective teacher behaviors and the

distinctive features of language teachers provide a good basis for designing the new form. The quantitative findings of the study showed that the final version of the form was highly reliable, though many items were positively skewed, showing that they may not be ideal for distinguishing between teachers. The factor analysis of the form in this study revealed only a single factor, suggesting that students do not distinguish between various aspects of good teaching; this situation may be inferred as evidence for a halo effect in student's ratings.

Key Words: teacher evaluation forms, students' evaluation of teachers, effective teacher behaviors, distinctive feature of language teaching, designing a questionnaire, validity of forms.

ÖZET

YABANCI DİL OLARAK İNGİLİZCE ÖĞRETİMİ ALANINA MAHSUS ÖĞRENCİLER İÇİN GEÇERLİ BİR ÖĞRETMEN DEĞERLENDİRME FORMU TASARLANMASI

Ayça Özçınar

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi Programı

Tez Yöneticisi: Yrd. Doç. Dr. Jodee Walters

Temmuz 2011

Öğrenciler için Öğretmen Değerlendirme Formları 1920'lerden bu yana kullanılmaktadır. Pek çok araştırmacı o zamandan beri bu formların geçerliliğini tartışmaktadır. Pek çok çalışmada alana özgü öğretmen değerlendirme formu tasarlama girişimi olmuştur.

Bu form Yabancı Dil Olarak İngilizce öğretimi alanında öğrenciler için öğretmen değerlendirme formu hazırlamayı amaçlamaktadır. Kullanılan araçlar bir çevrimiçi anket, formun üç farklı hazırlama sürecindeki halleri ve formun kendisidir. Katılımcılar Anadolu Üniversitesi Yabancı Diller Yüksekokulu'ndan farklı dil yeterlilik seviyelerinden 392 öğrenci (başlangıç, alt-orta, orta, üst-orta ve yüksek seviyeden) ve 21 öğretmen ve farklı üniversitelerden 3 öğretmendir. Çevrimiçi anket ve formun versiyonları üzerine yapılan görüşmeler nicel olarak, formun son halinin pilot çalışması nitel olarak analiz edilmiştir.

Nitel Analiz göstermiştir ki şurada kullanılmakta bulunan Anadolu Üniversitesi Yabancı Diller Yüksekokulu öğretmen değerlendirme formu yetersizdir. Etkili öğretmen davranışları ve yabancı dil öğretmenlerinin ayırt edici özelliklerine ait literatür yeni bir form hazırlayabilmek için iyi bir kaynak olmuştur. Nitel analiz göstermiştir ki, formun son halinin güvenilirliği yüksektir, fakat pek çok maddenin öğrenci değerlendirmeleri pozitif yöne eğilimlidir, bu durum formun öğretmen değerlendirmesi için çok ideal olmadığı sonucunu ortaya çıkarmaktadır. Faktör analizi sonucu formun tek bir faktörden oluştuğu görülmektedir, bu durum öğrencilerin iyi öğretmenliğin farklı yönlerini ayırt etmediklerini işaret etmektedir ki bunu ağıl etki (halo effect) olarak yorumlayabiliriz.

Anahtar Kelimeler: öğretmen değerlendirme formları, öğrencilerin öğretmenleri değerlendirmesi, etkili öğretmen davranışları, dil öğretiminin ayırt edici özellikleri, bir anket tasarlama, formun geçerliliği.

ACKNOWLEDGEMENTS

I would like to show my gratitude to my thesis supervisor Asst. Prof. Dr. JoDee Walters for her guidance and feedback.

I would like to express my deepest gratitude to Asst. Prof. Dr. Philip Durrant who also supervises my thesis throughout the study. His encouraging and understanding attitude makes the thesis writing process better.

I would like to thank to Asst. Prof. Dr. Julie Mathews Aydınlı for her support and advice while choosing my thesis subject.

I am also indebted to many of my colleagues at Anadolu University for their support. This thesis wouldn't have been possible without the valuable contributions of Serkan Geridönmez, Özlem Kaya, Sercan Sağlam, Ayşe Dilek Keser, Figen Tezdiker, Meral Melek Ünver, Eylem Koral, Eda Kaypak, Başak Erol, Aycan Akyıldız Uygun, Gözde Şener, Kadir Özsoy, Çağdaş Delibaş, Revan Koral, Hilal Peker, Müge Kanatlar, Hülya Emeklioğlu İpek, Zuhal Kıyılık, and Zehra Herkmén Şahbaz.

I wish to thank to Bahar Tunçay, Zeynep Erşin, Ebru Gaganaş, Ebru Öztekin, Figen İyidoğan, Nihal Yapıcı Sarıkaya, and Elizabeth Pullen for their sincere friendship.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZET	v
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xi
LIST OF FIGURES	xii
CHAPTER I - INTRODUCTION	1
Background of the study	1
Statement of the problem	4
Research questions	5
Significance of the study	6
Conclusion	6
CHAPTER II: LITERATURE REVIEW	8
Introduction	8
Teacher evaluation forms for student ratings	8
Validity concerns	10
Construct validity	10
Discriminant validity	11
Convergent validity	11
Criterion validity	12
Consequential validity	12
Face validity	14
Content validity	15
Items that are not related to teaching performance	15

Distinctive features of language teachers	15
Effective teacher behaviors	18
Reliability.....	25
Designing a specific TEF for students' ratings	25
Conclusion	29
CHAPTER III: METHODOLOGY	30
Introduction.....	30
Setting	30
Participants.....	31
Instruments.....	33
Procedure	35
Data Analysis	38
Conclusion	38
CHAPTER IV: DATA ANALYSIS	39
Introduction.....	39
Results	40
Literature review & online form.....	40
The analysis of version 1.....	42
The analysis of version 2.....	51
The analysis of version 3.....	61
The analysis of the piloting of the final version	69
Participants.....	70
Distribution of the data.....	71
Reliability.....	72
Factor Analysis.....	72
The differences of the average scores in terms of language skills	73

Conclusion	76
CHAPTER V- CONCLUSION	77
Introduction.....	77
General results and discussion	78
Implications.....	85
Limitations of the study	86
Suggestions for further research.....	87
Conclusion	87
REFERENCES	89
APPENDIX A - FINAL VERSIONS OF THE FORM, ENGLISH AND TURKISH	93
APPENDIX B - AN EXAMPLE TRANSCRIPT OF THE INTERVIEWS.....	95
APPENDIX C - CURRENT TEF FOR STUDENTS' RATINGS AT AUSFL.....	97
APPENDIX D - PUBLISHED SOURCES FOR EACH QUALITY AND ITEM ...	98

LIST OF TABLES

1 - Teacher participants in the study at all stages	33
2 - Student participants in the study at all stages.....	33
3 - The age of the student participants in the piloting of the form	70
4 - Participants' distribution across proficiency level and language skill.....	70
5 - The distribution of responses of the students for each item.....	71
6 - The reliability scores of the final form	72
7 - The average scores of teachers in terms of skills	74

LIST OF FIGURES

1 - categories of effective teacher behaviors and characteristics and their references in the literature	20
2 - The initial version of the form.....	43
3 - Neglected subjects in the initial version of the form	50
4 - The second version of the form.....	52
5 - The third version of the form	62
6 - The final version of the form.....	69
7 - The five normally distributed items.....	72

CHAPTER I – INTRODUCTION

Teacher Evaluation Forms (TEFs) are forms that are completed by students to gather information about teaching practice in the classroom. They have been in use since the 1920s. The results of the ratings given by students are analyzed and interpreted both for summative and formative purposes. Summative evaluation is used to rank the quality and effectiveness of teaching; formative evaluation is used to improve the process (Cashin & Downey, as cited in Wolfer & Johnson, 2003, p112). There are two types of TEF: generic and field specific. A generic form consists of general items about teaching, whereas a specific form consists of specific items that may differ in different disciplines.

At Anadolu University, a generic TEF for students' ratings is used for formative purposes in all departments at the end of each term. On the basis of Burden's (2008) argument that there is a need for a specific form to evaluate language teachers' performance, this study aims to prepare a valid EFL- specific TEF for students' ratings to evaluate the performance of language teachers by considering the distinctive features of EFL.

Background of the study

TEFs for students' ratings are used as a source of data to provide information about teacher performance to improve education (Mace, 1997). The University of Washington administered TEFs for the first time in history to gather information from students about teachers' performance in the early 1920s. TEFs were used to meet students' demands for accountability and informed course selections in the 1960s, to improve faculties in the 1970s, and for administrative purposes in the 1980s and 1990s. Recently, TEFs have been used to improve education and to meet

the demand for accountability (Ory, as cited in Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier, & Moore, 2007).

Although TEFs are in wide use, there are still validity concerns regarding them. In terms of a form, validity means that it measures what it is supposed to measure. There are many validity types. Construct validity is one of the most controversial concerns about the validity of TEF for students' ratings. Construct validity is the ability of TEFs to measure effective teaching, which is an abstract concept. For example, if the communicative skills of teachers are evaluated in the form, the items should be based on observable behaviors that are able to measure the abstract concept of "being communicative".

Another validity concern about TEF for students' ratings is consequential validity, which shows whether the effects of the results of TEF for students' ratings have positive or negative consequences in education. Student ratings are commonly used in one of two ways: summatively or formatively. A summative evaluation is used to take personnel decisions by administrations, such as promotion. A formative evaluation is used to improve teachers' instruction. The positive and negative consequences of the use of these two types of evaluations have been discussed in the literature. Proponents of student ratings argue that students' ratings on teaching performance are invaluable sources which provide for teaching improvement and effectiveness (Panasuk & Lebaron, as cited in Nasser & Fresko, 2002). Improving instruction, promoting teacher and learner growth and reflection, and diagnosing future learning needs are argued to be the benefits of student ratings (Doyle, as cited in Nasser & Fresko, 2002). On the other hand, opponents of student ratings criticize the negative effects of ratings on the quality of education, such as teachers' tendency

to give high grades to students to gain high ratings from them or fear of the personnel decisions that administrations make based on the results of TEFs. According to them, student ratings are only a measure of teacher popularity and students are not capable of making reliable and valid judgments (Nasser & Fresko, 2002). For this reason, the opponents warn administrators about the danger of misuse of data while giving personnel decisions (Adams, 1997). Emery, Kramer and Tian (2003) suggest that if students' ratings can destroy a teacher's career, students may use their ratings as a threat against their teacher, so teachers may have tendency to give high grades, which decreases the quality of education. On the other hand, McKeachie (as cited in Greenwald, 1997) suggests that high grades do not automatically indicate a fear of judgment summatively by administrators. Yunker and Yunker (2003) mentioned that if a teacher is good at his job, he naturally has high ratings.

The content validity of TEF for students' ratings is another validity concern. If a form has content validity, it means that it measures what it is supposed to measure (Brown, 2004). Therefore, while preparing a new TEF, the researcher should pay attention to the content of the discipline that she is going to study. Neuman (2001) mentioned that there are differences between disciplines in terms of several factors, such as types of teaching, preparation time, practice, curriculum assessment of students, program review, teaching approaches, and teaching outcomes. Burden (2008) argues that EFL teachers have distinctive qualities from teachers of other disciplines and so should be evaluated differently. He also suggests that some items of TEFs for students' ratings are not about teachers' performance, but course books, syllabus, and so on, which are not under the control of teachers.

Lee (2010) criticized TEFs for not being sensitive to socio-cultural differences. He compared the views of Japanese EFL students in his own study and the views of similar European students studied by Borg (2006) about effective teacher qualities. European students suggested that being humorous, flexible and creative were qualities of effective teachers, whereas Japanese EFL students, who prefer more traditional attitudes in lecturing, do not agree with them on this point. In response to all these criticisms about TEF for students' ratings, the present study aims to design an EFL-specific TEF which measures language teachers' performance by considering Turkish students' attitudes towards language education.

Statement of the problem

There are many studies about the validity and reliability of TEF for students' ratings (Adams, 1997; Emery, Kramer & Tian, 2003; Greenwald, 1997; Kidd & Latif 2004; Nasser & Fresko, 2002), the interpretations of the results of the form (Damron, as cited in Emery, Kramer & Tian, 2003, p.4; Fresko, 2001) and the interpretation of the ratings on the Likert-scale (Block, 1998). Moreover, Borg (2006) and Lee (2010) explored the perceptions of students' about effective language teachers and Burden (2008), who also mentioned the need of an EFL-specific TEF for students' ratings, studied the distinctive features of language teachers. However, I am not aware of any research which has focused on designing a valid EFL-specific TEF for students' ratings.

At Anadolu University School of Foreign Languages (AUSFL), a generic online teacher evaluation form is used for all disciplines. Since it is a generic form, the present form does not consist of EFL-specific items to evaluate language teachers' performance. Moreover, although the form is used to evaluate teachers, a

number of items on the form (e.g. ‘The appropriateness of the course book and other materials to the objectives of the course’) are not under teachers’ control. Moreover, the present form consists of many ambiguous items that may be difficult to interpret to rate (e.g. ‘The lessons being conducted in an interesting way’). When considering these problems of the present form, there is an urgent need to design a valid EFL-specific TEF for students’ ratings.

Research questions

The purpose of this study was to design a valid EFL- specific teacher evaluation form (TEF) for student ratings which evaluates the performance of language teachers by considering the distinctive features of EFL. The main research question can be defined as:

- What items should be included in an EFL-specific teacher evaluation form for student ratings?

It is intended that in the course of answering this question, a number of other points will become clear. In particular:

- To what extent is the evaluation form currently in use at AUSFL satisfactory for using with EFL courses?
- To what extent are concepts from the literature on effective teacher behaviors a good basis for creating a teacher evaluation form?
- Is the construct of ‘good teaching’, as evaluated by students, a unitary construct or can it be divided into distinct sub-categories?
- Are there differences in the ways in which students evaluate teachers of different language skills (such that evaluation forms ought to be made skill-specific)?

Significance of the study

The study aims to contribute to the literature by meeting the need for a EFL-specific TEF for students' ratings (Burden, 2008). In addition to this, the study may provide an example for all disciplines which use generic forms to evaluate their teachers' performances, to prepare a specific TEF which considers the distinctive features of their discipline (Neuman, 2001).

At the local level, the current study aims to explore to what extent the evaluation form currently in use at AUSFL is satisfactory; to what extent concepts from the literature on effective teacher behaviors are a good basis for creating a teacher evaluation form; whether the construct of 'good teaching' as evaluated by students is a unitary construct or is divided into distinct sub-categories; and whether there are differences in the ways which students evaluate teachers of different language skills. At the end of the study, it is aimed to design a valid EFL-specific TEF for students' ratings to measure language teachers' performance at AUSFL. In addition, the form may help other language programs to be aware of the need for an EFL-specific TEF for students' ratings to evaluate language teachers' performance. In addition, language teachers may have the chance to see their weak and strong points in their instruction more clearly by looking at their results. The data also may provide information about the needs of institutions for in-service training.

Conclusion

The chapter presents the background of the study, the statement of purpose, the research questions, and the significance of the study. The next chapter will present the literature review of the study. The third chapter will present the

methodology of the study. The fourth chapter will present the data analysis; and the last chapter will present the conclusion of the study.

CHAPTER II: LITERATURE REVIEW

Introduction

This study aims to prepare a valid EFL-specific teacher evaluation form (TEF) for students' ratings to evaluate the performance of language teachers by considering the distinctive features of EFL.

In this chapter, research about evaluation forms, the validity and reliability concerns of the forms, the distinctive features of disciplines, and the inadequacy of general evaluation forms in evaluating the performance of field teachers will be presented.

Teacher evaluation forms for student ratings

TEF for students' ratings are instruments in which student ratings are used as a source of data to provide information about teacher performance to improve education (Mace, 1997). There are generic and specific TEF for students' ratings. A generic TEF for students' ratings consists of general items about effective teaching that can be used by all departments. However, a specific TEF for students' ratings is designed by considering the distinctive features of a department with its field related items.

TEF for students' ratings were first used at the University of Washington in the 1920s (Seldin, 1993). Since then, many institutions have used TEF for students' ratings to evaluate the effectiveness of their instructors, for both summative and formative purposes. Formative evaluation provides information for teachers to improve their instruction. Teachers learn their weak and strong points in their

instruction with the help of formative evaluation to teach more effectively.

Summative evaluation assesses the outcomes of teaching.

Formative evaluation provides information for instructional effectiveness, which leads teachers to improve the quality of teaching (Marsh, as cited in Nasser & Fresko, 2002). All assessing activities which give feedback are formative assessments, and they are used for modifying teaching (Black & Wiliam, as cited in Wei, 2010). According to these definitions, if teachers are evaluated for formative purposes, the information is used to adapt teaching and learning to meet student needs (Black & Wiliam, as cited in Boston, 2002). With the help of formative evaluation, teachers become aware of the weak points in their instruction and make necessary modifications to improve student success (Boston, 2002). However, summative evaluation, which focuses on the outcomes of teaching, is used for personnel decisions by administrators (Marsh, as cited in Nasser & Fresko, 2002). It can be defined as effectiveness evaluation. Ranks and scores have primary importance in summative evaluation. Summative evaluation aims to assess the effects or outcomes of teaching. Some studies point out that students may abuse their ratings as a threat against teachers (Emery, Kramer & Tian, 2003). Benz and Blatt suggested that if students are dissatisfied with their grades, their evaluations become negative (as cited in Adams, 1997). Therefore, using the data gathered by TEF for students' ratings for only summative purposes may affect teachers' performance negatively. Teachers may be afraid of losing their jobs because of the results of summative evaluation, so they may want to have good relations with students and their primary concern may become receiving high marks in TEF for students' ratings instead of teaching. Therefore, students may take control of the lessons by

threatening their teachers, so providing discipline in the classroom may become a very difficult issue for teachers. As a result of this, the quality of the lessons may decrease in the long run (Emery, Kramer & Tian, 2003).

Consequently, using TEF for formative purposes may provide information about teachers' performance to improve education quality, yet if it is used for summative purposes, the outcomes may be quite the opposite. No matter what the purpose is, there are still validity concerns about TEFs.

Validity concerns

If a tool measures what it is supposed to measure, it can be considered valid (Golafshani, 2003). Many studies consider TEF for students' ratings as the most valid source of data on evaluating teaching effectiveness (McKeachie, 1997) because it provides information about improving instruction, promoting teacher and learner growth and reflection, and diagnosing future learning needs (Doyle, as cited in Nasser & Fresko, 2002). However, there are many concerns about the validity of TEF for students' ratings.

Construct validity

One of the most controversial concerns about the validity of TEF for students' ratings is construct validity. A construct is an attempt to describe an abstract concept that may not be measured directly (Brown, 2004). For example, motivation is a psychological construct that cannot be measured directly. Therefore, some observable behaviors which indicate motivation status should be identified to measure whether people are motivated. The object of construct validity is to explore whether the constructs of a subject matter are evaluated in an instrument (Brown, 2004). In the present study, the effectiveness of a language teacher is a construct, so

the items of TEF for students' ratings should serve to measure the construct. There are two types of construct validity, discriminant and convergent. If an instrument has low correlations with unrelated constructs, it has discriminant validity. If an instrument has high correlations with other indicators of a subject matter, it has convergent validity.

Discriminant validity

The discriminant validity of an instrument shows the influence of unrelated factors on subject matter (Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier & Moore, 2007). For example, the effect of students' easy grade expectations on student ratings is a factor unrelated to effective teaching. According to Madden, Dillon and Leak (2010), student evaluations of teaching are biased because of students' grade expectations, so students' easy grade expectations cause increases and decreases to some extent in the results of TEF for students' ratings. Therefore, teachers who give high grades may receive high scores in TEF for students' ratings. However, Yunker and Yunker (2003) claimed that if a teacher is good at his job, students learn more and get better grades, so student ratings are possibly higher. Therefore, a correlation between high grades and high ratings may not be evidence of invalidity all the time (McKeachie, as cited in Greenwald, 1997).

Convergent validity

Convergent validity of an instrument shows a correspondence between similar indicators that measure the same thing (Onwuegbuzie et al., 2007). There are many other indicators, such as peer feedback, classroom observation and self-evaluation forms, which are used to evaluate teaching. If various indicators' results have similarities with each other, it means that each has convergent validity.

Therefore, while designing a valid TEF for students' ratings, other indicators may help to show the validity of the instrument.

Criterion validity

Similar to convergent validity, criterion validity also compares an instrument with another indicator; the only difference is that the indicator is an instrument that has previously been proven valid (Hughes, 2004). Criterion validity has two sub-categories. Predictive validity shows whether there is a correlation between an instrument's results and the score of another instrument in the future (Hughes, 2004). Concurrent validity differs from predictive validity in terms of administration time. In concurrent validity, the two instruments are administered at the same time.

Consequential validity

Consequential validity, which shows whether the results of TEF for students' ratings are used to improve education beneficially, is another controversial point. First, the results of summative evaluations of teachers by administrators for personnel decisions may affect consequential validity. Teachers' judgments in the classroom may be affected by the fear of penalty, so getting on well with students may become their primary goal. Therefore, the results cause a decrease in the quality of education instead of improvement. Madden, Dillon and Leak (2010) pointed out the halo effect in student ratings, and suggested that student ratings are only a measure of teacher popularity and students are not capable of making reliable and valid judgments. Second, the difficulty of interpreting the results may cause misinterpretation or overinterpretation and it may not provide beneficial information to improve education. For consequential validity, Damron suggested that just having a valid questionnaire is not a solution, and that using the results in a beneficial way

to improve education needs accurate interpretation (as cited in Emery, Kramer & Tian, 2003). Marsh and Roche suggested professional consultancy for teachers to interpret the results, to avoid misinterpretations (as cited in Greenwald, 1997). Many instructors have difficulty in interpreting the results of TEF for students' ratings to make appropriate adjustments in their practice even if they know statistics (Avisar, Bar-Zohar, Shiloach, as cited in Nasser & Fresko, 2001). Nasser and Fresko (2001) and Block (1998) attempted to explore interpretation problems derived from TEF for students' ratings results.

Nasser and Fresko (2001) conducted a study on developing course evaluation instruments and creating a consultation group to help teachers to understand the results of the instrument. They worked on the interpretations of the results of questionnaires with the participation of teachers of different fields. They found that most of the participants claimed that consultation was very helpful to understand the student ratings. However, those who had positive attitudes towards consultation were teachers who believed in the validity of student ratings at the beginning of the consultation process. Although the researchers claimed that consultation make only a few differences in practice, they suggested that in the long run consultation might be beneficial to understand the results, so the data might be used to improve the quality of education.

Block (1998) conducted a study on the interpretation of questionnaire items from an end-of-course evaluation form with 24 students in Barcelona. He asked to what extent participants interpret questionnaire items similarly and to what extent they understand the same meaning in the numbers on a one to five rating scale. Block gave the data of three items, which were about overall evaluation of a teacher,

making class interesting, and punctuality. Using interviews, he analyzed the interpretations of respondents to each question. First, he found that students were confused whether the evaluation was about only the teacher or about the class in general, although the form mentioned that it was only about the teacher. The researcher suggested that the word “overall” caused that confusion. In addition, the individual differences of respondents played a great role in the results. For example, some respondents simply did not believe in perfection and their highest rank was 4. Therefore, there was a variety of interpretations of an item. Second, similar to “overall”, the word “interesting” had different interpretations for students. Some of them related the word with activities, others focused on teachers’ characteristics. Third, students were confused about the term “punctuality”, so some of them interpreted it as student lateness. Some other students overgeneralized their attitudes towards a lesson, and gave high ranks if they liked the lesson in general. In conclusion, the study shows that the rating scales and the items of the questionnaire may have various interpretations, and the researcher suggested that these interpretations should be explored while designing a questionnaire, so the data gathered by the questionnaire can be used beneficially.

Face validity

Face validity gives information about the perceptions of the participants in the evaluation process about whether the instrument used in the study looks valid (Onwuegbuzie et al, 2007). Students are the respondents of TEF for students’ ratings, so asking their perceptions on the validity of TEF for students’ ratings is a simple way to explore face validity.

Content validity

Content validity shows whether an instrument measures what it is supposed to measure (Brown, 2004). For example, if you want to measure the knowledge of simple present tense, but give a test about present perfect tense, the test lacks content validity. If the aim is to assess teaching performance, the items should only measure teaching performance. There are a number of possible threats to the content validity of TEFs.

Items that are not related to teaching performance

One possible threat is that TEF for students' ratings may consist of some items that do not evaluate teachers' performances. Burden (2008) conducted a study on the perceptions of sixteen ELT teachers on the usefulness of the questions of a TEF for students' ratings in Japan. He found that the TEF for students' ratings had some items for evaluating the syllabus, students' self-evaluation, the course book and supplementary material, and classroom equipment, in addition to items for teaching performance. The participants suggested that some items are not related to teachers' ability to teach, so the results are about not only teachers' performance but also various different aspects of an EFL classroom.

Distinctive features of language teachers

Another possible threat to the validity of the form is disregarding the distinctive features of language teaching. Although there are some common points with other disciplines in terms of teaching, teaching language requires some distinctive features. Burden (2008) points out, the distinctiveness of language teaching causes content validity problems for teacher evaluations which are not taken into account. Neuman (2001) studied disciplinary differences in university

teaching, and he mentioned that disciplines may differ from each other in terms of several factors, such as types of teaching, preparation time, practice, curriculum, assessment of students, program review, teaching approaches, and teaching outcomes. Neuman's article raises the question of how language teachers are different from teachers of other subjects.

Borg (2006) conducted a study on the distinctive characteristics of foreign language teachers. He suggested that disciplinary characteristics and good language teacher behaviors give evidence for the distinctiveness of language teaching. In Borg's study, various participants - including postgraduate students of TESOL, language teacher conference delegates, subject specialists (chemistry, mathematics, science, and history teachers), Hungarian pre-service teachers of English, and Slovene undergraduates in English – gave their opinions about the distinctiveness of language teachers. According to the findings of the study, first of all, only language teachers use the subject matter as a medium to teach, even if students do not understand at first, so the nature of the subject matter is distinctive. This may create a distance between students and a language teacher. In addition, Borg suggested that language is a constantly changing matter; therefore, language teaching should be innovative. Moreover, during the process, students have opportunities to see practical outcomes of the education, unlike in other fields. Furthermore, the content is more complex, containing not only grammar knowledge but also communicative skills. Therefore, teachers need to organize their instruction to provide interaction such as group and pair work, to provide communication and student participation in the process. Borg pointed out that language teachers need to prepare extracurricular activities to provide naturalistic environments that provide authenticity, more than in

other subjects. As a result of communication based activities, teachers and students have closer relations than in other disciplines. Having field knowledge, ability to organize, motivate and explain the subject, and being fair and helpful to students are claimed to be some characteristics of a language teacher. During these interactive activities, students should not be afraid of making mistakes or participating, so errors are acceptable during the learning process, unlike in other fields. Borg also suggested that only in language learning, non-native professionals are compared with native speakers who have no education in language teaching. Finally, there are various goals of learners; therefore, language learning may be characterized by these goals. For example, the method of teaching English for academic purposes or for travelling might be different from each other. Borg's findings can be considered as evidence of the need for a specific TEF for students' ratings for EFL teachers to evaluate their effectiveness fairly, by considering the distinctiveness of language teaching.

In the light of Borg's study (2006), Lee (2010) conducted research on the uniqueness of language teachers in Japan with 163 college level EFL students. Lee summarized the research findings under four main headings, which are the nature of the subject matter, content, teaching approach, and teacher personality. Borg suggested that the language learning process has more practical relevance to real life. However, in Lee's study, participants did not agree with this idea because the nature of language teaching differs in Japan. Most of the participants are the students of engineering departments, so the researcher suggested that they may consider that science and engineering courses are more practical in their future plans. The participants mentioned that the content is also different. Japan's new policy on the importance of the knowledge of culture, in addition to grammar learning, and the

introduction of TOEIC, which is an exam that Japanese students have to take at the end of each academic year, broaden the scope of language learning content.

Therefore, language teachers should focus on developing grammar, reading, writing, speaking, listening, and the culture of a new language in Japan, unlike teachers of other subjects. In terms of teaching approach, language teachers use communicative approaches to develop interactive skills, and extend student participation in the classroom. They also avoid correcting errors so as not to discourage student participation. Unlike in Borg's study, in Lee's study language teachers were not characterized as more humorous, creative or flexible. However, positive attitudes and enthusiasm were also shown as two characteristics that distinguish language teachers from other teachers, as it was mentioned in Borg's study. All in all, the findings of this study show that although there is a common assumption that language teachers have distinctive features, these differences can change according to socio-educational contexts.

With these studies in mind, the distinctive features of language teachers will be kept in mind while designing an EFL-specific evaluation form for students' ratings to provide information about the needs of language learning in addition to some common points of effective teacher behaviours.

Effective teacher behaviors

Another factor to consider in ensuring content validity is defining what counts as effective teacher behavior (ETB). There is no universal accepted definition of ETB because there are several dimensions that embody ETB (Johnson & Ryan as cited in Devlin & Samarawickrema, 2010). It is possible to say that this is because most of the studies in the literature focus on only some of the dimensions and

disregard the others. For example, Spencer and Schmelkin (2002) focus on personal characteristics of effective teachers, such as caring for students, showing respect to students' thoughts, and clarity in communication. Ramsden (as cited in Çakmak, 2009), on the other hand, mainly focuses on pedagogical skills, such as modifying teaching strategies according to the particular subject matter, students and teaching environment, encouraging critical thinking and problem solving skills, demonstrating an ability to transform and extend knowledge, setting clear goals and using appropriate assessment methods, and providing high quality feedback to students. While these studies have focused on a single aspect of teaching, Bailey argues that teaching effectiveness is not a simple construct, but a synthesis of many factors such as teachers' personal characteristics, content knowledge, caring behavior and the culture of the teaching environment (as cited in Rahimi & Nabilou , 2010).

In the present study, the researcher use the information gathered from the general and EFL-specific ETB literature, as well as the literature about TEF for students' ratings, while creating an item pool. Figure 1 (beginning on page 20) presents the categories of effective teacher behaviors and characteristics as described in the literature and their references in the literature. The items are grouped into four categories based on the information gathered from the literature. Each category consists of four to 15 items.

Figure 1 - categories of effective teacher behaviors and characteristics and their references in the literature

Quality in the literature	Reference
Personal Characteristics	Spencer & Schmelkin, as cited in Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier & Moore (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie et al. (2007) Marlin & Niss (1980)
Being understanding / empathy	Marlin & Niss (1980), Alhija & Fresko (2009)
Caring for students	Okpala & Ellis, as cited in Onwuegbuzie et al. (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie et al. (2007) Marsh (1984) Shishavan & Sadeghi (2009) (http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm)
Communicator	Spencer & Schmelkin, as cited in Onwuegbuzie et al. (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie et al. (2007) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Cruickshank, Bainer and Metcalf, as cited in Şahenk (2010)
Providing a peaceful environment	(http://celt.ust.hk/instr/instr11.htm) Alhija & Fresko (2009) Jang, Guan & Hsieh (2009) Şahenk (2010) Shishavan & Sadeghi (2009)
Friendly	Marsh (1984) Cruickshank, Bainer and Metcalf, as cited in Şahenk (2010) Shishavan & Sadeghi (2009)
Having positive attitude	Cruickshank, Bainer and Metcalf, as cited in Şahenk (2010)
Patient	(http://www.edpsycinteractive.org/materials/tchlrmnd.html) Shishavan & Sadeghi (2009)

Field Knowledge	Kane, Sandretto &Health, as cited in Onwuegbuzie et al. (2007) Content Knowledge - Okpala &Ellis, as cited in Onwuegbuzie et al. (2007) Knowledge of Subject- Emery, Kramer & Tian (2003)
Having field knowledge	Buskit, as cited in Onwuegbuzie et al. (2007) Cruickshank, Bainer and Metcalf, as cited in Şahenk (2010) Madden, Dillon & Leak (2010) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Jang, Guan & Hsieh (2009)
Having knowledge in other disciplines	Bell (2005)
Having knowledge about target languages' culture	Shishavan & Sadeghi (2009) Bell (2005)
Having knowledge about target languages' culture	Bell (2005)
Making explanations	(http://celt.ust.hk/instr/instr11.htm).
Answering students' questions	(http://www.questionpro.com/akira/showSurveyLibrary.do?surveyID=88&mode=1)
Improving himself in his profession	Cruickshank, Bainer and Metcalf, as cited in Şahenk (2010) Shishavan & Sadeghi (2009)
Giving examples	Greilmen-Furhman &Geyer as cited in Onwuegbuzie et al. (2007) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Jang, Guan & Hsieh (2009) (http://www.edpsycinteractive.org/materials/tchlrmnd.html)
L1+L2	Bell (2005)
Pedagogical Skills	Kane, Sandretto &Health as cited in Onwuegbuzie, et al. (2007) Presentation skills- Crumbley, Henry, and Kratchman as cited in Onwuegbuzie, et al. (2007) Teaching Skills- Okpala &Ellis as cited in Onwuegbuzie, et al. (2007)

Being Prepared	Crombley, Henry & Kratchman as cited in Onwuegbuzie et al. (2007) Emery, Kramer & Tian (2003) (http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm) Madden, Dillon & Leak (2010) McGrath, Yeung, Comfort & McMillan (2005) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Marsh (1984)
Stating objectives	(http://www.sussex.ac.uk/tldu/ideas/eval/ceq) McGrath, Yeung, Comfort & McMillan (2005) Marsh (1984) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Jang, Guan & Hsieh (2009) Madden, Dillon & Leak (2010) Shishavan & Sadeghi (2009)
Passing on his knowledge to students	Was added to the form based the information gathered from online form
Answering questions	(http://www.questionpro.com/akira/showSurveyLibrary.do?surveyID=88&mode=1) Wennerstorm & Heiser (1992) Jang, Guan & Hsieh (2009) (http://www.edpsycinteractive.org/materials/tchlrmnd.html)
Having clear explanations	Greilmen-Furhman & Geyer, as cited in Onwuegbuzie et al. (2007) Marsh (1984) Wennerstorm & Heiser (1992) Alhija & Fresko (2009) (http://www.edpsycinteractive.org/materials/tchlrmnd.html)

Using course material appropriately	Use the course material in an interesting way
Teaching in an interesting way	Buskit , as cited in Onwuegbuzie et al. (2007) Wennerstorm & Heiser (1992) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Alhjiya & Fresko (2009) Jang, Guan & Hsieh (2009) Bell (2005)
Motivating students	McGrath , Yeung, Comfort & McMillan (2005) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) Shishavan & Sadeghi (2009)
Encouraging participating	(http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm) McGrath , Yeung, Comfort & McMillan (2005) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Alhjiya & Fresko (2009) Wennerstorm & Heiser (1992) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf)
Teaching Studying independently out of classroom	McGrath , Yeung, Comfort & McMillan (2005) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) Bell (2005)
Providing sufficient practice	Observable behavior for being prepared
Making real-life connections	Kane, Sandretto & Health, as cited in Onwuegbuzie et al. (2007) McGrath , Yeung, Comfort & McMillan (2005) (http://www.edpsycinteractive.org/materials/tchlrm.d.html)
Using relevant activities	Bell (2005)
Considering students proficiency level	Shishavan & Sadeghi (2009) Bell (2005)
Guiding students to study out of classroom	McGrath , Yeung, Comfort & McMillan (2005) Shishavan & Sadeghi (2009) Bell (2005)

Feedback & Assessment	A part of pedagogical skills category, yet the researcher decided to add a new category for evaluating feedback and assessment qualities.
Assessing what it has been taught	Shishavan & Sadeghi (2009)
Giving explanatory /useful feedback	http://www.sussex.ac.uk/tldu/ideas/eval/ceq McGrath , Yeung, Comfort & McMillan (2005) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) http://www.edpsycinteractive.org/materials/tchlrmtd.html Bell (2005)
Giving constructive feedback	http://www.edpsycinteractive.org/materials/tchlrmtd.html
Giving feedback in time	http://celt.ust.hk/instr/instr11.htm http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf Madden, Dillon & Leak (2010) http://www.edpsycinteractive.org/materials/tchlrmtd.html
Being fair while assessing	Crumbley, Henry, and Kratchman, as cited in Onwuegbuzie et al. (2007) Buskit, as cited in Onwuegbuzie et al. (2007) Emery, Kramer & Tian (2003) http://celt.ust.hk/instr/instr11.htm http://www.edpsycinteractive.org/materials/tchlrmtd.html http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf
Giving helpful feedback	http://www.sussex.ac.uk/tldu/ideas/eval/ceq http://celt.ust.hk/instr/instr11.htm McGrath , Yeung, Comfort & McMillan (2005) Madden, Dillon & Leak (2010)

Reliability

Aside from the aspects of validity that are mentioned above, reliability is another important aspect of preparing a good questionnaire. Whereas validity is concerned with whether an assessment measures what it is intended to measure, reliability is concerned with the consistency of assessment (Hughes, 1989). For example, if a questionnaire is administered on Monday or Wednesday, the scores may change although the rest of the variables can be controlled. Yet, the aim of research is to construct, administer and score instruments that provide similar scores even if administered at different times. So, similar scores are the evidence of the reliability of an instrument. Some of Hughes' suggestions for preparing a reliable test are taking enough behavior samples, not allowing participants too much freedom while filling a questionnaire, as occurs with open-ended questions, writing unambiguous items, providing clear and explicit instructions, paying attention to format, such as using bold when necessary, and providing a uniform and non-distracting environment for administration.

Designing a specific TEF for students' ratings

The present study's purpose is to design a valid and reliable specific TEF for students' ratings for EFL. The features of a valid and reliable instrument are mentioned above. In addition to these features, the study will focus on how to design a new instrument. The following section will review the literature about constructing a questionnaire.

Dörnyei (2003) defined a questionnaire as a written instrument that contains a series of questions, which can be classification, behavioral, and attitudinal questions, to gather data from respondents. Classification questions can be about demographic

characteristics, residential location, and level of education. Behavioral questions may ask about past actions, life-styles, habits or personal history of the participants.

Attitudinal questions cover attitudes, opinions, beliefs, interest, and values.

Dörnyei (2003) explains how to construct a questionnaire in detail. The first thing a researcher should do is to create an item pool by conducting interviews, and borrowing questions from the established literature. In addition, while designing a questionnaire, item wording plays an important role. The researcher should avoid using adjectives, universals (never, none, all), modifying words (only, just), loaded words (democratic, modern, free), negative constructions, and double barreled questions (two questions in one item). Furthermore, Dörnyei suggests that the order of questions is important. The researcher should choose opening questions carefully to take the attention of the participants. Classification questions should be asked at the end of the questionnaire to avoid creating resistance in the respondents. Adding open-ended questions at the end of a questionnaire is another suggestion to avoid possible negative consequences. For example, participants may get bored and spend less time to fill the questionnaire if they answer the open-ended questions' part first. Moreover, Dörnyei claims that conducting a pilot study is the most important priority of designing a questionnaire. Sudman and Bradburn suggested that "if you do not have the resources to pilot-test your questionnaire, don't do the study" (as cited in Dörnyei, 2003). Piloting a study helps to identify ambiguous wording, items that are difficult to reply to, overlapped and irrelevant items, problems with administration, scoring and processing, arranging the necessary length of time, and neglected subjects of the content. A researcher can modify his questionnaire through a pilot study.

A number of studies have been conducted on designing an instrument to evaluate teachers' effectiveness. Taylor, Reeves, Mears, Keast, Binns, and Ewings (2001) conducted research on the development and validation of a questionnaire to evaluate the effectiveness of evidence-based practice teaching (EBP) in the literature with 152 health care professionals. The items were chosen by reviewing the literature and borrowing from an established EBP questionnaire. To create a clear and short questionnaire, they interviewed 20 health care professionals, who gave feedback on both content and format, and reduced the number of items and modified the questionnaire. The results of the study showed that the instrument had moderate to high levels internal consistency, discriminative validity and responsiveness. So, the researchers concluded that the questionnaire was valid for measuring the impact of EBP training on participants' knowledge and attitudes toward EBP.

Another study was conducted by Dondanville (2005), who developed an instrument to assess effective teacher behaviors in athletic training clinical education. The participants were 145 students of an athletic training education program. The researcher created items by reviewing the relevant health literature, and prepared 20 items. Items were grouped in four categories. information, evaluation, critical thinking and physical presence. Then an expert panel of seven athletic training education program directors and clinical coordinators analyzed the instrument. The questionnaire was modified through the report of the expert panel. Finally, the researcher conducted a pilot test with a convenience sample of students to evaluate the reliability of the instrument. Students rated both their current and an ideal clinical instructor. He calculated the internal consistency of the items by looking at the scores of each item between these two ratings.

In clinical education, McGrath, Yeung, Comfort and Mc Millan (2005) developed a questionnaire to evaluate clinical dental teachers. One hundred and forty-eight dental students assessed 29 clinical dental teachers with the questionnaire in the study. The researchers created an item pool by reviewing the literature, gathering feedback from faculty staff, and organizing group discussions with dental students. First, they used both the student ratings of clarity and relevance and factor analysis to choose the items to put in the questionnaire. Then, after administration of the questionnaire, face validity, construct validity, criterion validity, and reliability were assessed. They asked the students to also rate the instructors globally, from very good to very poor, and they compared those ratings with the results of the questionnaire, for criterion validity. The researchers suggested that if the students' ratings of individual clinical dental teachers were similar and if there was minimal difference (not significantly different from zero) between the mean scores of two randomly allocated groups, the questionnaire could be considered to have construct validity. In the study, the difference between the groups was minimal or small, at less than 0.3; therefore, the form was considered as having construct validity.

Jang, Guan and Hsieh (2009) developed an instrument for assessing college students' perceptions of teachers' pedagogical content knowledge (PCK). They constructed their questionnaire by using the categories in Shulman's (1987) PCK. They tested the questionnaire in a pilot study with 16 novice teachers and 182 college students. Then they held interviews with several teachers and considered the suggestions of The Advancing Teachers' Teaching Excellence Committee to identify the overlaps and neglected subjects of the questionnaire. After modifying the instrument through the data gathered from the pilot study, the questionnaire was

administered. The researchers assessed the content and construct validity, and reliability of the questionnaire by analyzing 172 responses.

All these studies have some common points. The researchers created an item pool based on the data gathered from the relevant literature. The stakeholders were mostly professionals in the field of the study and students. The opinions of stakeholders about the items were taken in the interviews or group discussions. Then the data gathered from the interviews was used to modify the instrument. Finally, the content validity, the construct validity and reliability of the all forms was measured. Therefore, these studies create an outline for the present study to design a new questionnaire.

Conclusion

Assessing an abstract phenomenon such as teachers' performance is a complex issue that requires considering various aspects. The distinctiveness of a field, the key points of designing a questionnaire, and considering the validity and reliability of the instrument are the keystones of designing a new instrument. Burden (2008) pointed out that EFL classrooms need a specific form to evaluate the performance of language teachers. The study described in the following chapters will describe the preparation of a valid specific TEF for students' ratings for EFL by considering the previous studies on designing an instrument, the concerns of validity and reliability, and the distinctiveness of language teaching and effective teacher behaviors.

CHAPTER III: METHODOLOGY

Introduction

The purpose of this study was to design a valid EFL- specific teacher evaluation form (TEF) for student ratings which evaluates the performance of language teachers by considering the distinctive features of EFL. The main research question can be defined as:

- What items should be included in an EFL-specific teacher evaluation form for student ratings?

It is intended that in the course of answering this question, a number of other points will become clear. In particular:

- To what extent is the evaluation form currently in use at AUSFL satisfactory?
- To what extent are concepts from the literature on effective teacher behaviors a good basis for creating a teacher evaluation form?
- Is the construct of ‘good teaching’, as evaluated by students, a unitary construct or can it be divided into distinct sub-categories?
- Are there differences in the ways in which students evaluate teachers of different language skills (such that evaluation forms ought to be made skill-specific)?

Setting

The study was carried out at Anadolu University School of Foreign Languages (AUSFL), Eskisehir, Turkey. AUSFL provides compulsory intensive English language education for students of most departments in the university.

Therefore, each year, there are a great number of students who take a proficiency test at the beginning of their university lives. On the same day, students are given a placement test. Students who fail in the proficiency exam are placed in an appropriate level: Beginner, Elementary, Lower-intermediate, Intermediate, Upper-intermediate, or Advanced. AUSFL gives skill-based education and students at all levels take Grammar, Reading, Writing and Listening/Speaking courses.

Anadolu University uses a general online student evaluation form which has the same items for all departments. Students can rate their teachers' performance at the end of each term voluntarily. If at least twelve students rate a teacher's performance he/she can see the results of this evaluation.

Participants

Five groups participated in the study: First, fifteen language instructors of AUSFL¹ and three language instructors of other universities who were MA TEFL students at Bilkent University participated in an informal online questionnaire. They were recruited from among the researcher's colleagues and fellow MA TEFL students. The aim of including both AUSFL teachers and MA TEFL students, who were EFL teachers at different institutions, was to find out common topics that were considered important by English instructors from different institutions. Second, four EFL teachers from AUSFL were interviewed on their reactions to an initial item pool list. In this part, the participants had MA or PhD degrees or were experienced in preparing questionnaires, so their comments on both content and form of the questionnaire made a good first step. Third, five EFL teachers from AUSFL were interviewed about their reactions to the second draft of the possible item list. The

¹ 6 of 15 participants in the online questionnaire took part in the further interviews.

purpose of these interviews was to identify whether there were ambiguous or overlapping items, or neglected subjects. Fourth, ten students (two from each level - elementary, lower-intermediate, intermediate, upper-intermediate and advanced) and six AUSFL teachers participated in interviews which were based on the third version of the possible item list. In this part, having two students from each level aimed to make sure there was balance of views from across levels. The information gathered from these interviews, as in the previous set of interviews, was used to find out whether there were ambiguous or overlapping items, or neglected subjects. Finally, 34 classes from different levels filled out the final version of the evaluation form to evaluate their teachers. The classes were chosen randomly from a list of classes for each level.

All teachers who participated in the study were chosen from willing colleagues. Students were recruited by participating teachers and some other colleagues. In the piloting of the study, the researcher paid attention to balancing the number of students from different proficiency levels, yet, since piloting of the form was administered near the end of the semester, there were only a few students at the school because many high level students had already passed the preparatory class by using their grades on some other proficiency exams, like KPDS, TOELF and TOEIC, that are accepted by AUSFL.

Table 1 and Table 2 present background information about the teachers and students who participated in the study at all stages.

Table 1 - Teacher participants in the study at all stages

Gender	Female	Male		
	19	5		
Years of experience	1-4	5-8	9-12	13 +
	4	10	4	6
Graduate Programs	ELT	Literature	Translation	
	21	2	1	
Degree Programs	BA	MA	PhD	
	14	7	3	

Table 2 - Student participants in the study at all stages

Gender	Female	Male		
	218	165		
Age	17-19	20-21	22 +	
	212	148	23	
Level	Elementary	Low-Int	Int	Up-Int Advanced
	108	124	137	9 5

Instruments

An informal online questionnaire, a sample form with its three versions and the evaluation form itself were used in the study. After reviewing the literature, an informal online open-ended questionnaire was given to colleagues who are language instructors at AUSFL and MA TEFL students at Bilkent University. The researcher used the online questionnaire to gather information as quickly as possible because the information would be used as the basis of possible items in addition to the literature review study. An open-ended questionnaire type was used to avoid limiting participants' thoughts.

The questionnaire asked four questions:

1. If students evaluated your performance, what qualities would you expect them to be aware of?
2. Is language teaching different from teaching other disciplines? If yes, what are the differences?

3. Do language teachers have distinctive qualities from the teachers of other fields? If yes, what are the differences?
4. What are the mistakes that are included in the student evaluation forms?

The second instrument was an item pool, which had four versions, formed on the basis of a series of questionnaires and interviews with teachers and students and also the review of the literature. The development of each version of the item pool and the final version of the form itself is described briefly below (in the *Procedure* section) and in detail in Chapter 4 (see Appendix A to see the final version of the form in Turkish and English).

The researcher chose interviews while designing the evaluation form for a number of reasons. Kvale (1996) suggested that the interviewer has the opportunity to probe or ask follow-up questions, so interviews let researchers ask more complex questions than other types of data-collection methods. In the study, semi-structured face-to-face individual interviews, in which a list of the possible items of the TEF for student ratings was interpreted and discussed by all participants, were employed. In this way, the researcher had the chance to explore the participants' responses in detail. Using semi-structured interviews also helped the researcher categorize and compare the data easily in analyzing the process (see Appendix B to see the example of the transcript both in English and Turkish).

All instruments were in Turkish for many reasons, although they were translated to English for the purposes of inclusion in the thesis. The EFL- specific teacher evaluation form for student ratings would be in Turkish and would be completed by students at different proficiency levels at a Turkish university. The translation of an item is considered as a new item (G. Şerbetçioğlu, personal

communication, 27 March, 2011); therefore, preparing the items or conducting the interviews in English cannot provide the information that the study needs. In addition, both the teachers and students in the study were able to express themselves better in their mother tongue. It was felt that the more information about the items the researcher could gather, the more validity the study would provide.

Procedure

While designing a valid EFL-specific TEF for student ratings, the researcher followed the following procedure. First, an item pool was created by reviewing the literature of general student evaluation forms and distinctive features of language teaching. In addition to the information gathered from the literature review, an informal online open-ended questionnaire was prepared to identify the distinctive features of EFL teachers and it was given to the language instructors of AUSFL and MA TEFL students at Bilkent University. The informal online questionnaire asked teachers which of their qualities should be recognized by students, whether language teaching is different from teaching in other disciplines, whether language teachers have distinctive qualities from teachers of other fields, and what the mistakes are of current teacher evaluation forms that are used for student ratings.

According to the information gathered in the literature review and the online informal questionnaire, a list of possible items and categories was determined and the first draft of an item pool for EFL-specific TEF for student ratings was prepared. The first set of interviews consisted of four interviews with experienced teachers from AUSFL. The focus point was on both content and form of the possible items. The teachers interpreted the items and discussed their ideas about the items and

categories to build a better form with the researcher. The interviews on the first draft took 45 to 90 minutes each.

Throughout the study the interviews started with an explanation about the aim of the interviews, which was to find out whether there were ambiguous, overlapping items, neglected subjects or items that were not appropriate for students' evaluation of teachers. Then the interviewees' opinions about each item were asked one by one. The researcher asked what the interviewees understood from the item, whether the item was relevant and important in language learning, and whether they had any further comments on the item. See Appendix B for an example transcript of the interviews.

According to the information gathered in the interviews, some overlaps and neglected subjects were identified and the form was re-designed. The second version of the form, which consisted of some sub-categories under the main categories in the initial version, was examined with five other EFL teachers from AUSFL. Among the teachers in the second set of interviews, there were two experienced teachers, who had MA TEFL degrees, and three novice teachers. The aim of interviewing teachers who had different experience levels was to identify possible differences or similarities of thoughts about the items in a teacher evaluation form. According to the information gathered from the second set of interviews, some overlapping items and some items that were not appropriate for students' evaluation of teachers were eliminated; some neglected items were identified and added to the form; some ambiguous items were clarified; and some items that were too specific were generalized. By considering the new information gathered from the second set of interviews, the third version of the evaluation form was designed and the third set of

interviews was conducted with ten AUSFL students at five different proficiency levels (elementary, low-intermediate, intermediate, up-intermediate and advanced) and six other teachers. The purpose of these interviews was to explore whether there were items that were not appropriate for students' evaluation of teachers, ambiguous, overlapping, needed changes in language or should be combined. The second and third set of interviews took 25 to 35 minutes each. The series of interviews with the teachers and the students was examined to make adjustments to the form and to give it a final shape.

Finally, the last version of the evaluation form was piloted at AUSFL with 34 classes (382 students in total). The last version of the form consisted of 21 items. The form was designed as a series of Likert-type items. Students were informed that the evaluation form was a study in process and that it did not take the place of the form they already had online at the university web-site. Students were asked to complete the questionnaire to evaluate their teacher of the lesson in which the questionnaire was being administered. The lessons of the teachers were one of the four different skills (reading, writing, grammar, and speaking and listening) in language education. Each student participated in the study once and each teacher was evaluated no more than once. The researcher herself administered the questionnaire to provide clear instructions and to avoid the possible effects of different instructions. The teachers who were evaluated were not in the classroom during the administration in order to avoid possible effects of the teachers' presence on the students' evaluations. The data gathered from the Likert scale were analyzed on SPSS to find the internal consistency of the form.

Data Analysis

The study was based on qualitative analysis of semi-structured interviews on the versions of the form and the quantitative analysis of the piloting of the form at AUSFL. The initial item pool was designed on the basis of the information gathered from literature review and the online questionnaire. The second to fourth versions of the TEF for student ratings were designed based on the information gathered from the interviews in which participants discussed the previous version of the form. The analysis of all interviews focused on the common criticisms that participants made and all criticisms were supported by giving examples. The fifth version of the form was analyzed and information gathered about the distribution of the data, reliability of the form, factor analysis and whether there were differences in the average scores of some items in terms of language skills.

Conclusion

In this chapter, the information about the setting, participants, instruments, data collection procedures were presented, and data analysis procedures were briefly explained. In the next chapter, the data analysis procedure will be explained in detail.

CHAPTER IV: DATA ANALYSIS

Introduction

The purpose of this study was to design a valid EFL-specific teacher evaluation form for student ratings by considering the distinctive features of EFL. The research aimed to provide a basis for field-specific teacher evaluation forms in language teaching.

Qualitative data, the basis of the teacher evaluation form, was collected through an online questionnaire and semi-structured interviews. Quantitative data, which will be used to evaluate the form's reliability, were gathered in the form of responses to the evaluation form itself. This study gathered data from: 382 preparatory school students from different proficiency levels (110 elementary, 126 lower intermediate, 138 intermediate, 11 upper intermediate and 7 advanced level students); 21 instructors at Anadolu University School of Foreign Languages (AUSFL) (Eskişehir); and 3 instructors who are language instructors at other universities in Turkey and are students from the 2010-11 Bilkent University MA TEFL program.

In this chapter, the literature review and the qualitative analysis of the online form, the semi structured interviews on the initial, second and third version of the form, and the quantitative analysis of the piloting of the evaluation form will be presented in turn. The first section presents the information gathered from the literature review and an online questionnaire, which elicited the responses of 15 language teachers about teachers' expectations from students' awareness, the distinctive qualities of language teaching and language teachers and the problematic points of present teacher evaluation forms, to create an initial item pool. The second

to fourth sections present analyses of a series of semi-structured interviews with colleagues and students at AUSFL, each interview set addressing a new version of the form. The fifth section covers analysis of quantitative data gathered from the piloting of the evaluation form in which 382 students from different proficiency levels of AUSFL participated.

Results

Literature review & online form

Since there was no known specific evaluation form for evaluating language teachers' performance, an initial pool of items was created based on the literature on program evaluation, course evaluation, and general teacher evaluation forms and on an online questionnaire, administered to 15 teachers at AUSFL. The online questionnaire consisted of four open-ended questions and aimed to gather information about language instructors' beliefs and expectations about:

- students' awareness of teachers' performance;
- the distinctive qualities of language teaching;
- the distinctive qualities of language teachers;
- problematic items of the current AUSFL teacher evaluation form.

Fifteen of the 30 instructors surveyed responded. The first three questions of the form provided information about their beliefs about what qualities are important and distinctive for language teachers. The information gathered from these questions was used to choose appropriate qualities for language teaching from the great number of qualities that were found in the literature review. The last question provided information about teachers' views on the current TEF for students' ratings at AUSFL. Twelve teachers from AUSFL found the form unsatisfactory and made

two basic types of criticism of the current form (see Appendix C for the current TEF for students' ratings at AUSFL). First, they stated that it included items that were not under the teachers' authority. Four items, in particular, were criticized for not being under the teachers' control. These items measure the appropriateness of the course book and other materials to the objectives of the course, the appropriateness of homework, projects and other activities to the objectives of the course, the appropriateness of exams to the subjects of the course, and the ability of exams to evaluate students' success. All participants suggested that these items were not related to teachers' performance in the classroom because teachers have no right to choose the course book or other materials for the sake of standardization in education at the institution; and homework, projects and exams were prepared by the skill and level coordinators and the administration. Therefore, the teachers felt that it was not fair to judge teachers' performance in terms of the course book or material choice. While designing the form, the researcher therefore paid attention to avoid using irrelevant items and focused on creating items that measured teachers' performance particularly. The second criticism was that certain items were ambiguous. The participants suggested that the ambiguity of the items may lead students to misinterpretation. For example, it is strictly forbidden to be late to lessons, and teachers are not allowed to tolerate late comers. However, item 14 asks about the teacher's tolerance towards students. Many students accused teachers of not being tolerant in late coming situations. Therefore, the participants did not want to be evaluated by this item. The extent of the expected tolerance was not clearly stated so the possibility of misinterpretation of the item increased. With this in mind, the

researcher paid attention to clarify the purpose of the items clearly while creating them to avoid misinterpretations.

The initial form was designed considering the data gathered from the literature review and the online questionnaire. First of all, the categories that were commonly mentioned in the literature were identified: personal characteristics, field knowledge, pedagogical knowledge and feedback and assessment. Then, a large initial item pool was designed based on the information gathered from the literature review. Second, by analyzing the item pool and the online questionnaire, some frequently mentioned qualities were identified. These qualities were each assigned to a category. Finally, items aiming to evaluate each quality were written by the researcher. The published sources for each quality and item are shown in Appendix D.

The analysis of version 1

The information gathered from the literature review and the online questionnaire was used to design the initial pool of items for the teacher evaluation form. The first version of the form had four main categories, with between six and 16 questions appearing under each heading, 42 items in total (see Figure 2, page 43).

At this stage of the study, the researcher interviewed four colleagues at AUSFL, all of whom had MA or PhD degrees in teaching English as a foreign language. The interviews, which took 45 to 90 minutes, were tape-recorded and transcribed by the researcher. The interview language was Turkish.

Figure 2 - The initial version of the form

Categories	Items
Personal Characteristics	<ol style="list-style-type: none"> 1. Our teacher behaves in an understanding way towards us. 2. Our teacher is interested in my problems. 3. Our teacher is able to communicate with us efficiently. 4. Our teacher provides a peaceful environment in the classroom. 5. Our teacher behaves like a friend towards us. 6. Our teacher has a positive attitude towards us. 7. Our teacher is patient while correcting our mistakes.
Field Knowledge	<ol style="list-style-type: none"> 1. Our teacher has enough field knowledge. 2. Our teacher's knowledge in other disciplines affects language learning positively. 3. Our teacher teaches not only language but also its culture. 4. Our teacher gives importance to teaching language with its culture. 5. Our teacher is able to make necessary explanations while instructing. 6. Our teacher has the knowledge to answer our questions about the lesson. 7. Our teacher improves herself all the time in her field. 8. Our teacher is able to give necessary examples which are relevant to the lesson. 9. Our teacher can explain the differences and similarities between Turkish and English Language.
Pedagogical Skills	<ol style="list-style-type: none"> 1. Our teacher comes to class prepared. 2. Our teacher presents the objectives of the lesson clearly. 3. Since the objectives of the lesson are clear, my learning process becomes meaningful. 4. Our teacher is able to pass on her knowledge to us. 5. Our teacher answers my questions about the lesson. 6. Our teachers' explanations are clear. 7. Our teacher uses the course materials well. 8. Our teacher teaches in an interesting way. 9. Our teacher motivates us to learn English. 10. Our teacher encourages us to participate in the activities. 11. Our teacher teaches us to study independently outside of the classroom. 12. Our teacher gives us sufficient practice. 13. Our teacher uses activities that are like real-life situations. 14. Our teacher uses activities that encourage using what we have learned. 15. Our teacher explains the lesson in a way that is suitable to our proficiency level. 16. Our teacher guides us to study outside of the classroom.
Feedback & Assessment	<ol style="list-style-type: none"> 1. Our teacher assesses what we have learned in the lesson by quizzes. 2. Our teacher gives explanatory feedback on our homework. 3. Our teacher corrects our mistakes without making us feel bad. 4. Our teacher gives feedback on time. 5. Our teacher is fair while evaluating our in-class performance. 6. Our teacher's feedback helps me to improve myself.

Interviewees made four basic types of criticism of the first version of the form: that items were ambiguous or overlapping, that certain important subjects were neglected and that certain subjects were not appropriate for students' evaluation of teachers. In the following paragraphs, these criticisms will be described in turn.

The most commonly-cited problem of the initial evaluation form was ambiguous items. The different interpretations of the items and direct statements about items being ambiguous made some changes essential on the form. According to the data gathered from the interviews, unclear objectives, abstract items, language use and wording caused ambiguity on the initial version of the form.

First, the participants noted that some items were unclear about their objectives and so may be misinterpreted unless clarified. For example, Item 1 in the personal characteristics category was “Our teacher behaves in an understanding way towards us”. All participants pointed out that this could be interpreted as being tolerant towards students who do not do their homework, of students’ psychological problems, or being patient while instructing. Three participants noted that the purpose of being understanding should have been stated in the sentence. In response to such comments, in the second version of the form, two items were added in which the purpose of being understanding was mentioned clearly:

- Our teacher patiently answers our questions relevant to the lesson. (Item 2b)
- Our teacher patiently helps us to solve our problems when we do not understand the subject. (Item 2c)

Another example of ambiguity was Item 8 in the pedagogical skills category, “Our teacher teaches in an interesting way”. Three participants suggested that if the purpose of the item was to motivate students, this should have been stated explicitly. For this reason, the item was not changed but put under the subcategory of motivation in the second version of the form so that its phrasing could be discussed further.

Last, in the feedback and assessment category, Item 6, “Our teacher’s feedback helps me to improve myself” was interpreted as asking for an evaluation for the students rather than of the teachers. One of the participants suggested that if the purpose of the item was to learn whether teachers gave useful, explanatory, or clear feedback, it should have been given in the sentence. With this in mind, three items were added under the subcategory of giving clear feedback in the second version of the form:

- Our teacher gives clear feedback on homework and projects etc.(Item 18a)
- Our teacher explains our mistakes on exams in addition to our grades.
(Item 18b)
- Our teacher explains the correct answers of the exercises we have done.
(Item 18c)

Using abstract expressions in the items was another most commonly-cited ambiguity source on the first version of the form. For example, the phrase “providing a peaceful environment” in Item 4 of the personal characteristics category was found too abstract by all participants, who suggested some observable behavior to overcome the ambiguity problem caused by this abstract expression. The interviewees suggested adding items about avoiding arguments and stress, providing a classroom environment in which people feel free to express their thoughts, providing an environment that motivates students, being fair towards every student, and encouraging students’ participation in lessons. In response to such comments, in the second version of the form, some observable items were added. For example:

The items about avoiding arguments and stress that were added to the second version of the form were:

- Our teacher provides enough discipline in the classroom. (Item 10a)
- Our teacher helps us to handle the difficulties that we face. (Item 2a)
- Our teacher helps us to overcome the problems that we face during language learning. (Item 20b)

The items about providing a classroom environment in which people feel free to express their thoughts that were added to the second version of the form were:

- Our teacher listens to us. (Item 1c)
- Our teacher respects our thoughts. (Item 1d)

The items about providing an environment that motivates students that were added to the second version of the form were:

- Our teacher makes lessons interesting by using various materials. (Item 16b)
- Our teacher makes course material interesting. (Item 15b)
- Our teacher is able to draw our attention to the lesson by using extra materials. (Item 16c)
- Our teacher prepares visual, audio or written materials in addition to the course book to make lessons more enjoyable. (Item 15d)
- Our teacher uses enjoyable activities while revising the subjects. (Item 15f)

The items about being fair towards every student that were added to the second version of the form were:

- Our teacher takes care of each student equally. (Item 3a)
- Our teacher gives us equal right to speak. (Item 3b)
- Our teacher assesses us fairly. (Item 21a)
- Our teacher gives equal time to giving feedback to each student. (Item 21b)

The item about encouraging students' participation in lessons that was added to the second version of the form was:

- Our teacher is able to give us courage to participate in lessons. (Item 16c)

Another example of abstract statements in the items was having field knowledge (Item 1 in the field knowledge part). One of the participants suggested that this quality could be evaluated by an observable behavior which was passing on one's knowledge to the students. With this in mind, a subcategory called "passing on knowledge" was added to the pedagogical skills category of the second version of the form. The items under this category were:

- Our teacher teaches clearly. (Item 12a)
- Our teacher answers our questions related with the subject clearly. (Item 12b)
- Our teacher is ready to deal with the problems we may face relevant to the subject. (Item 12c)
- Our teacher considers our language proficiency level while giving the lesson. (Item 12d)

Lastly, Item 1, which was about being prepared for lessons in the pedagogical skills category, was found abstract by two participants, who suggested that this item could be evaluated by some observable behavior, such as preparing materials, answering exercises and questions of students, and giving examples. The following items were in response to this:

- Our teacher prepares extra material when necessary. (Item 13b)
- Our teacher explains the correct answers of the exercises we have done. (Item 18c)

- Our teacher gives clear examples when necessary to make the subject more clear. (Item 7d)

Another commonly-cited ambiguity on the first version of the form was language use and wording. The interviews were held in Turkish; therefore, the items that were discussed during the interviews were criticised in terms of their language and wording in order to clarify the meanings. Although some comments were based on the Turkish language, some criticisms also affected the English version of the items. Here some examples will be given of the latter group. Item 2, “Our teacher presents the objectives of the lesson clearly” and Item 3, “Since the objectives of the lesson are clear, my learning process becomes meaningful” in the pedagogical skills category were good examples of ambiguity caused by wording. All participants suggested that “the objectives of lesson” was difficult to be interpreted by students who did not know the terms that were used in pedagogy. In response to such comments, some observable items were added in addition to this item to be discussed in the second version of the form:

- Our teacher clearly explains why we learn subjects. (Item 11a)
- Our teacher informs us what we will learn in the lesson. (Item 11c)

Item 3, “Our teacher corrects our mistakes without making us feel bad” in the feedback and assessment category was another example of ambiguous language use. One of the participants suggested that instead of the phrase “without making us feel bad”, a clearer alternative would be “Our teacher gives constructive feedback”. With this in mind, two items were added under the subcategory of attitude while giving feedback in the second version of the form:

- Our teacher corrects our mistakes in a positive manner. (Item 20a)

- Our teacher expresses our mistakes in an appropriate manner. (Item 20c)

The second most cited problem of the initial evaluation form was overlapping items. During discussions, it was noticed that some qualities were described in different items under different categories. To avoid repetition of evaluating the same quality many times, at first all participants suggested getting rid of overlapping items. However, at the end of the interviews, they suggested instead identifying subcategories for each main category and extending the pool of items. This caused more overlapping in the short-term but enabled a wider range of items to be evaluated. For example, it was suggested that Item 11, “Our teacher teaches us to study independently outside of the classroom” and Item 16, “Our teacher guides us to study outside of the classroom” in the pedagogical skills category overlapped according to all participants, who preferred Item 11 because they believed that Item 11 brought learner autonomy in teaching. Two participants noted the importance of providing learner autonomy in language learning. With this in mind, a subcategory was added to the second version of the form to find out the most appropriate item to evaluate this quality. The following items were added in response to this:

- Our teacher supports our out-of-class learning so that language learning is not restricted to the classroom. (Item 17a)
- Our teacher shows us the way to do self-study outside of the classroom. (Item 17b)

The third most cited problem of the initial evaluation form was neglected subjects. The participants suggested several subjects which ought to be added to the form. These subjects, and the items which were written to address them, are summarized in Figure 3.

Figure 3 - Neglected subjects in the initial version of the form

Category	Suggested Subjects for Items	In response to suggestions, items that were added to the second version of the form
Personal Characteristics	<ol style="list-style-type: none"> 1. being cheerful 2. having a sense of humour 	<ul style="list-style-type: none"> • Our teacher has a sense of humour. • Our teacher cheers up the classroom when necessary. • Our teacher tolerates our jokes.
	<ol style="list-style-type: none"> 3. being empathetic 	<ul style="list-style-type: none"> • Our teacher empathizes with us.
	<ol style="list-style-type: none"> 4. being innovative (e.g. using technology while instructing) 	<ul style="list-style-type: none"> • Our teacher follows an innovative method while instructing. • Our teacher uses innovations in language teaching in her classroom. • Our teacher is open to innovations in instructing. • Our teacher follows innovations in her field.
	<ol style="list-style-type: none"> 5. being aware of students' individual differences and learning styles while instructing 	<ul style="list-style-type: none"> • Our teacher is aware of our individual differences. • Our teacher prepares visual, audio or written materials in addition to the course book to make lessons more enjoyable.
Field Knowledge	<ol style="list-style-type: none"> 1. having knowledge of other fields 2. having liberal knowledge 	<ul style="list-style-type: none"> • Our teacher has enough liberal culture to support language education. • Our teacher is a role model for us with her liberal culture.
Pedagogical Skills	<ol style="list-style-type: none"> 1. using intonation, mimics, body language to strengthen instructing 	<ul style="list-style-type: none"> • Our teacher uses her intonation to make her expressions stronger. • Our teacher uses her body language and mimics to make her expressions stronger.
Feedback & Assessment	<ol style="list-style-type: none"> 1. rewarding efforts and success 	<ul style="list-style-type: none"> • Our teacher uses expressions like “Well-done” etc. to reward us.

The final type of problem identified in the initial evaluation form was that some items asked for evaluations which students are not likely to be able to provide. For example, one of the participants suggested that students do not have enough information to make assumptions on the professional improvement of teachers. Therefore, it was suggested that Item 7, “Our teacher improves herself all the time in her field” in the field knowledge category be omitted.

As a result of the analysis of the interviews about the initial version of the form, the items of the second version of the form were designed. In the following section, the analysis of the interviews about the second version of the form will be described.

The analysis of version 2

As described above, the information gathered from the analysis of the first set of interviews with four colleagues was used to design a second version of the item pool. The second version had four main categories, with between two and nine sub-categories appearing under each heading, 77 items in total. (see Figure 4, beginning on page 52).

At this stage of the study, the researcher interviewed five other colleagues at AUSFL. The interviews, which took 30 to 35 minutes, were tape-recorded and transcribed by the researcher. The interview language was Turkish.

Interviewees made four basic types of criticism of the second version of the form. Some criticisms were similar in type to those made of the first form, others were new. In order of frequency, the main criticisms were that items overlapped, were ambiguous, were too specific, neglected important points or were not appropriate for students' evaluation of teachers. In the following paragraphs these criticisms will be described in turn.

In response to the criticisms by the participants of the initial interviewees, the item pool was extended by adding sub-categories. As a result of this, the most commonly-cited problem of the second version of the evaluation form was overlapping items. During discussions, it was noticed that some qualities were described in different items under different sub-categories. As was described in the

previous section, a certain number of overlapping items were intentionally included in the form. Throughout the interviews, the main purpose was to find out the best expression among overlapping items and omit the rest of them. Some examples will be described in the following paragraphs.

Figure 4 - The second version of the form

Personal Characteristics	
1. Communicator	<ul style="list-style-type: none"> a. Our teacher is a good communicator. b. Our teacher communicates with us well. c. Our teacher listens to us. d. Our teacher respects our thoughts. e. Our teacher expresses himself clearly. f. Our teacher uses her intonation to make her expressions stronger. g. Our teacher uses her body language and mimics to make her expressions stronger. h. Our teacher is aware of our individual differences. i. Our teacher makes us feel that he values us. j. Our teacher empathizes with us.
2. Patient	<ul style="list-style-type: none"> a. Our teacher helps us to handle the difficulties that we face. b. Our teacher patiently answers our questions relevant to lessons. c. Our teacher patiently helps us to solve our problems when we do not understand the subject.
3. Fair	<ul style="list-style-type: none"> a. Our teacher takes care of each student equally. b. Our teacher gives us equal right to speak.
4. Having sense of humour	<ul style="list-style-type: none"> a. Our teacher has a sense of humour. b. Our teacher cheers up the classroom when necessary. c. Our teacher tolerates our jokes.
5. Having liberal culture	<ul style="list-style-type: none"> a. Our teacher has enough liberal culture to support language education. b. Our teacher is a role model for us with her liberal culture.
6. Innovative	<ul style="list-style-type: none"> a. Our teacher follows an innovative method while instructing. b. Our teacher uses innovations in language teaching in her classroom. c. Our teacher is open to innovations in instructing.
Field Knowledge	
7. Having field Knowledge	<ul style="list-style-type: none"> a. Our teacher has enough knowledge about her field b. Our teacher answers our questions related with the subject clearly. c. Our teacher teaches clearly. d. Our teacher gives clear examples when necessary to make the subject more clear. e. Our teacher has enough knowledge to explain cultural features of the target language. f. Our teacher uses the target language fluently. g. Our teacher uses the target language accurately.
8. Innovative	<ul style="list-style-type: none"> a. Our teacher follows innovations in her field.

Pedagogical Skills	
9. Being prepared	<p>a. Our teacher comes to lessons well prepared.</p> <p>b. Our teacher does planned and scheduled lessons.</p> <p>c. Our teacher uses activities relevant to our subject.</p> <p>d. The activities that our teacher uses complement each other.</p>
10. Providing discipline	<p>a. Our teacher provides enough discipline in the classroom.</p> <p>b. Our teacher is able to provide a peaceful classroom environment to do lessons.</p>
11. Stating objectives of the lessons	<p>a. Our teacher explains why we learn subjects clearly.</p> <p>b. Our teacher explains the objectives of lessons clearly.</p> <p>c. Our teacher informs us what we will learn in the lesson.</p>
12. Being able to pass knowledge	<p>a. Our teacher teaches clearly.</p> <p>b. Our teacher answers our questions related with the subject clearly.</p> <p>c. Our teacher is ready to deal with the problems we may face relevant to the subject.</p> <p>d. Our teacher considers our language proficiency level while giving the lesson.</p>
13. Using reinforcement	<p>a. Our teacher does enough practice and repetition to help us understand the subjects easily.</p> <p>b. Our teacher prepares extra material when necessary.</p>
14. Time management	<p>a. Our teacher uses lesson time efficiently.</p> <p>b. Our teacher gives enough time to complete tasks.</p>
15. Using Materials	<p>a. Our teacher uses course materials efficiently.</p> <p>b. Our teacher makes course material interesting.</p> <p>c. Our teacher is able to draw our attention to the lesson by using extra materials.</p> <p>d. Our teacher prepares visual, audio or written materials in addition to the course book to make lessons more enjoyable.</p> <p>e. Our teacher uses technological features of the classroom.</p> <p>f. Our teacher uses enjoyable activities while revising the subjects.</p> <p>g. Our teacher is able to support the content of the course book when necessary.</p>
16. Motivating students	<p>a. Our teacher teaches in an interesting way.</p> <p>b. Our teacher makes lessons interesting by using various materials.</p> <p>c. Our teacher is able to give us courage to participate in lessons.</p> <p>d. Our teacher plays an active role to make us conscious of the reasons for language learning.</p> <p>e. Our teacher enlightens us to be aware of how beneficial language learning will be in our future life.</p> <p>f. Our teacher enlightens us how useful language learning is in real life.</p> <p>g. Our teacher creates realistic situations to make language learning meaningful.</p>
17. Providing learner autonomy	<p>a. Our teacher supports our out-of-class learning so that language learning is not restricted to the classroom.</p> <p>b. Our teacher shows us the way to do self-study outside of the classroom.</p>

Feedback & Assessment	
18. Giving clear feedback	a. Our teacher gives clear feedback on homework and projects etc. b. Our teacher explains our mistakes on exams in addition to our grades. c. Our teacher explains the correct answers of the exercises we have done.
19. Timing	a. Our teacher gives feedback on time. b. Our teacher checks our homework regularly.
20. Having a constructive attitude	a. Our teacher corrects our mistakes in a positive manner. b. Our teacher helps us to overcome the problems that we face during language learning. c. Our teacher expresses our mistakes in an appropriate manner.
21. Fair	a. Our teacher assesses fairly. b. Our teacher gives equal time to giving feedback to each student.
22. Evaluating what has been taught	a. Our teacher evaluates what he has taught in the exams that he prepares. b. Our teacher gives homework relevant to our subjects that we have in the lessons.
23. Rewarding students	a. Our teacher makes constructive comments to improve our language learning. b. Our teacher uses expressions like “Well-done” etc. to reward us.

To begin with, in the personal characteristics category, Items 1c, “Our teacher listens to us”, 1d, “Our teacher respects our thoughts” and 1i, “Our teacher makes us feel that he values us”, which were all under the sub-category of being a communicator, overlapped. One of the participants mentioned that if someone respected one’s thought, it showed that he listened to her thoughts. All participants criticized 1i for not being clear and they interpreted the item as showing respect to one’s thoughts as in Item 1d. Therefore, Items 1c and 1i were omitted and only Item 1c (Item 1) was left in the third version of the form.

Another example was that Items 7c in the field knowledge category and 12c in the pedagogical skills category were exactly the same item: “Our teacher teaches clearly”. It was pointed out by two participants that this item also overlapped with Item 1e, “Our teacher expresses himself clearly” in the personal characteristics

category. Items 1e and 12c were therefore omitted in the third version of the form and only Item 7c was left in the third version of the form (as Item 2).

Lastly, all participants suggested that Item 21b, “Our teacher gives equal time to giving feedback to each student” in the feedback and assessment category overlapped with Item 3a, “Our teacher takes care of each student equally” in the personal characteristics category. With this in mind, Item 21b was omitted in the third version of the form and the wording of Item 3a was changed to “Our teacher takes care of us as much as we need” (Item 14 in the third version) because the participants noted that teacher did not have enough time to be interested in each student individually in crowded classes.

The second most frequently mentioned criticism was the ambiguity of the items in the second version of the form. The differences among interpretations of the items and the criticisms of the participants about items showed that there were still ambiguous items on the form.

To begin with, all participants noted that Item 1h, “Our teacher is aware of our individual differences” in the personal characteristics category was ambiguous because it was not clear how it contributed to language learning. To clarify the aim of the item, two alternatives were suggested as additions to the third version of the form.

- Our teacher teaches by considering different learning styles in the classroom.
- Our teacher is aware of our interest areas.

However these alternatives were not found appropriate by the researcher according to the criticisms of the other similar items on the form. The first alternative

included the phrase “learning styles”, a pedagogical term that students were not able to understand. In addition to this, the second item needed a clarification about its purpose. For this reason, the item was modified, “Our teacher makes lessons interesting by considering our interests” (Item 12 in the third version).

Item 15a, “Our teacher uses course materials efficiently”, in the pedagogical skills category, was another example of an ambiguous item. All participants suggested adding some examples of course materials in the sentence. With this in mind, a new item, “Our teacher uses the course book, board, technological equipment in the classroom efficiently” (Item 22 in the third version), was added.

Lastly, some items were found too abstract to be evaluated. For example, four participants noted that Items 1a, “Our teacher is a good communicator” and 1b, “Our teacher communicates with us well” were ambiguous items because they were not observable. They suggested eliminating the item. With this in mind, the items were omitted in the third version of the form.

Based on the information gathered regarding the first version of the form, the researcher had attempted to extend the item pool by adding new sub-categories and items. This attempt caused some over specification of the items in the second version of the form. Throughout the interviews, the participants were expected to choose the best expression among some items or to help the researcher find more general expressions to evaluate desired qualities of teachers.

For example, evaluation of the quality of being patient during the teaching process was expressed in two items in the personal characteristics category.

2b. Our teacher patiently answers our questions relevant to lessons.

2c. Our teacher patiently helps us to solve our problems when we do not understand the subject.

Three participants suggested adding a general item about patience instead of these two items, which were too specific. They noted that a teacher should be patient all through the process. Therefore, a new item, “Our teacher behaves patiently during language teaching”, which was a more general statement, was added to the third version of the form (Item 13, in the third version).

Another example was Items 16d, 16e, 16f and 16g, which were about making language learning meaningful for students in the pedagogical skills category.

16d. Our teacher plays an active role to make us conscious of the reasons of language learning.

16e. Our teacher enlightens us to be aware of how beneficial language learning will be in our future life.

16f. Our teacher enlightens us how useful language learning is in real life.

16g. Our teacher creates real like situations to make language learning meaningful.

Four participants preferred Item 16g and one participant found 16f more clear; therefore, items 16d and 16e were omitted in the third version of the form. Both 16g and 16f were put on the form for further discussion. The wording of Item 16g was changed to make it more general “Our teacher makes language learning meaningful”.

Items 14a, 14b, 19a and 19b were all about time management. However items 14b, 19a and 19b were found too specific by all participants, who suggested that Item 14a covers all of the others.

14a. Our teacher uses lesson time efficiently.

14b. Our teacher gives enough time to complete tasks.

19a. Our teacher gives feedback on time.

19b. Our teacher checks our homework regularly.

With this in mind, Item 14a was put on the third version of the form (Item 16, in the third version). The rest of the items were eliminated.

In addition to overlapping, ambiguous and over specific items, the participants noted that there was one neglected subject on the form, which was about the professional attitude of teachers in lessons. Two participants suggested adding “Our teacher does not bring her personal problems to the classroom” as a sign of teachers’ professional attitude. With this in mind, this item was added to the third version of the form (Item 9, in the third version). Another participant suggested adding “Our teacher does not take problems in the classroom personally” as a sign of providing discipline and handling problems that occurred in the classroom in a professional way. In response to these comments, a new item, “Our teacher handles disciplinary problems that occur in the classroom” (Item 21) was designed for the third version of the form.

The last commonly mentioned criticism was that some items were not appropriate for students’ evaluation of teachers. First, the participants believed that some items were not about teachers’ performance in the classroom but about their personality directly. For example, all participants noted that Item 4a, “Our teacher has a sense of humour” in the personal characteristics category should be omitted because students were not able to evaluate teachers’ sense of humour. One of the participants believed that she did not have to entertain students during language

teaching and if the item was used, it gave students the right to expect teachers to behave as showmen. Two participants noted if the aim of the item was motivating students in a positive way by using a sense of humour, it should be expressed directly. In response to these criticisms, a new item “Our teacher cheers up the classroom when necessary” and two alternatives “Our teacher uses her sense of humour to motivate us in the lesson.” and “Our teacher makes lessons cheerful/enjoyable to improve our motivation.” were added in the third version of the form (Item 11, in the third version).

In addition, some items included some pedagogical terms and concepts which students could not be expected to know about. For example, two of the participants suggested that students were not able to evaluate Item 6a, “Our teacher follows an innovative method while instructing” because they did not have enough knowledge about methods. Moreover, two of the other participants noted that students could interpret the item as using technology, authentic materials or materials relevant to current issues instead of professional innovations. Therefore, the item was not used in the third version of the form. Next, Items 7g, “Our teacher uses the target language fluently” and 7f, “Our teacher uses the target language accurately” required a high language proficiency level to evaluate. Therefore, three participants believed that students were not able to evaluate teachers’ fluency and accuracy in language accurately. They suggested that students were able to make a general comment based on their observations in the classroom; therefore, a general statement about teachers’ language proficiency level could be added to the form. With this in mind, these two items were omitted and a new item, “Our teacher uses the target language well” (Item 7) was put on the third version of the form.

The last criticism by participants was about two items that were intended to evaluate learner autonomy. According to the interpretations of the participants of items 17a, “Our teacher supports our out-of-class learning so that language learning is not restricted to the classroom” and 17b, “Our teacher shows us the way to do self-study outside of the classroom”, these items were about encouraging self-study instead of learner autonomy. They suggested a number of alternatives to evaluate providing learner autonomy:

- Our teacher makes us conscious about being aware of our responsibility in our language learning.
- Our teacher encourages/guides us to take part in our own learning process actively.
- Our teacher provides choices during lessons to give us a chance to take part in decision making in our own education.

With this in mind, instead of adding new items in the third version of the evaluation form, learner autonomy was discussed by the participants to gather more information on the subject. Another controversial subject was evaluating the liberal culture of teachers in the forms. This subject was also discussed by the participants to gather more information on the subject.

As a result of the analysis of the interviews about the second version of the form, the items of the third version of the form were designed. In the following section, the analysis of the interviews about the third version of the form will be described.

The analysis of version 3

The information gathered from the analysis of the second version of the evaluation form was used to design the third pool of items for the teacher evaluation form. In the previous version of the form, using subcategories caused many overlapping problems. In addition to this, when subheadings were used, the aims of the items were announced to the participants which may affect their interpretations of the items. To avoid overlapping and such effects on interpretations, the subcategories of the second version were omitted in the third version of the form, which had 31 items in total (see Figure 5 on page 62). In addition to 31 items, four alternative items from the previous versions of the form or that appeared during the interviews were discussed with the participants in Items 11, 17, 20 and 23.

At this stage of the study, there were two groups of participants. The first group comprised six colleagues and the second group comprised two students from each proficiency level, elementary, lower-intermediate, intermediate, upper-intermediate and advanced levels (10 students in total). The interviews, which took 25 to 35 minutes each, were tape-recorded and transcribed by the researcher. The interview language was Turkish.

Interviewees made five basic types of criticism of the third version of the form. Some criticisms were similar in type to those made of the first and second form, others were new. In order of frequency, the main criticisms were that items were ambiguous, overlapped, were not appropriate for students' evaluation of teachers, needed changes in language, and should be combined. In the following paragraphs these criticisms will be described in turn.

Figure 5 - The third version of the form

Items
1. Our teacher respects our thoughts.
2. Our teacher teaches clearly.
3. Our teacher answers our questions related with the subject clearly.
4. Our teacher considers our language proficiency level while giving the lesson.
5. Our teacher gives explanatory examples when necessary.
6. Our teacher does a sufficient number of exercises and repetition to help us to learn better.
7. Our teacher uses the target language well.
8. Our teacher uses her intonation and body language to make her expressions stronger.
9. Our teacher does not bring her personal problems to the classroom.
10. Our teacher is able to explain cultural features of the target language when necessary.
11. Our teacher cheers up the classroom when necessary
Alternatives for Item 11: Our teacher uses her sense of humour to motivate us in the lesson. Our teacher makes lessons cheerful/enjoyable to improve our motivation.
12. Our teacher makes lessons interesting by considering our interests.
13. Our teacher behaves patiently during language teaching
14. Our teacher takes care of us as much as we need.
15. Our teacher comes to lessons well prepared.
16. Our teacher uses lesson time efficiently.
17. Our teacher explains the objectives of the lessons in every lesson.
Alternative item for 17: Our teacher explains clearly why we learn subjects.
18. Our teacher creates a peaceful classroom environment to conduct the lesson
19. Our teacher creates a classroom environment in which everybody feels free to express themselves.
20. Our teacher handles disciplinary problems.
Alternative item for 20: Our teacher effectively handles problems that spoil the atmosphere of the lesson.
21. Our teacher uses the course book, board, technological equipment in the classroom etc. effectively. /efficiently
22. Our teacher gives us courage to participate in lessons.
23. Our teacher makes language learning meaningful.
Alternative item for 23: Our teacher explains how we can use what we have learned in real life.
24. Our teacher guides us in how to study outside of the classroom by ourselves.
25. Our teacher gives clear feedback on homework and projects etc.
26. Our teacher explains our mistakes on exams in addition to our grades.
27. Our teacher explains the correct answers of the exercises we have done.
28. Our teacher makes constructive criticisms to improve our language learning process.
29. Our teacher appreciates our efforts and success.
30. Learner Autonomy (the subject is discussed)
31. Liberal Culture (the subject is discussed)

The most commonly mentioned criticism was ambiguity. The main source of ambiguity was wording problems. For example, Item 4, “Our teacher considers our

language proficiency level while giving lessons”, aimed to evaluate whether teachers pay attention to students’ English proficiency levels while instructing in English. Half of the participants interpreted the item as asking whether the teacher used L1 while instructing. For example, three lower proficiency level students complained about teachers’ not using L1 while instructing and one higher proficiency level student complained about using L1 while instructing. Moreover, two teachers noted that they used L1 while instructing when necessary. To avoid misinterpretations, the wording of the item was changed on the final form to “Our teacher adjusts her English by considering our proficiency levels in her lesson” (Item 3 on the final version of the form).

Another example of an ambiguous item was Item 11, “Our teacher cheers up the classroom when necessary”. All teachers and students mentioned that the aim of the item was not clear. The item could be interpreted as playing games, watching movies or making jokes all the time. All teachers and six students suggested that if the aim of the item was motivating students in the classroom in a positive way, it should be mentioned directly in the item. For this reason, Item 11 was eliminated and a new item “Our teacher motivates us in the lesson positively” (Item 7 on the final version of the form) was added to the final version of the form. Moreover, one of the interpretations of this item was using pair and group tasks to provide variety in terms of activities in the lesson and to focus on communicative skills. With this in mind, a new item, “Our teacher effectively uses pair and group works in the lessons” (Item 13 on the final version of the form) was added.

Items which had alternative versions were also found ambiguous. For example, Item 17, “Our teacher explains the objectives of the lessons in every

lesson.” had an alternative item, “Our teacher clearly explains why we learn subjects”. Nine students and all teachers found the original item too abstract and the alternative item clearer. They mentioned that students may have difficulty in interpreting the phrase “the objectives” and the original form of the item was interpreted as announcing the objectives at the beginning of lessons, yet this can be done throughout the lessons by teachers. Therefore, the original version of the form was found unclear in terms of its wording. With this in mind, the alternative version of the item was added to the final version of the form (Item 12). Another example of an alternative item was Item 23, “Our teacher makes language learning meaningful” and its alternative item, “Our teacher explains how we can use what we have learned in real life”. Five teachers and all students found the alternative item clearer because they mentioned that making connections with real life and language learning motivated students. They suggested that the phrase “making language learning meaningful” was an abstract expression and it was not clear that what student should expect teachers to do to provide this meaningfulness in lessons. In response to these comments, the alternative item was added to the final version of the form (Item 17).

The second frequently mentioned criticism was overlapping items. For example, Item 1, “Our teacher respects our thoughts” was found to overlap with Item 3 (listening to and answering students’ questions) and Item 28 (being constructive while giving feedback). For this reason, Item 1 was omitted and “Our teacher uses appropriate and constructive language while commenting on our performance” (Item 20) was added in the final version of the form.

Another example of overlapping items were Item 18, “Our teacher creates a peaceful environment to conduct the lesson”, Item 19, “Our teacher creates a

classroom environment in which everybody feels free to express themselves”, and Item 20, “Our teacher handles disciplinary problems”, which were designed to evaluate the classroom management skills of teachers in the classroom. They were all put on the form to find the best expression for evaluating classroom management quality. First, Item 18 was interpreted as providing appropriate physical conditions for the lesson by four teachers and five students. Because of the fact that teachers do not have any authority on physical conditions, the item did not serve its aim and so was eliminated. Next, two teachers criticized Item 19 for being ambiguous because it was not clear to what extent and on what subjects that students should be free to express their thoughts. Nine students noted that too much freedom in the classroom may cause problems so teachers may lose the control of the classroom. According to the interpretations, the item did not serve its purpose. In response to these criticisms, Item 19 was eliminated. Lastly, an alternative item for Item 20 was provided on the third version. Five teachers and nine students preferred the alternative item, “Our teacher effectively handles problems that spoil the atmosphere of the lesson” because the phrase “disciplinary problems” was found too negative. As a result of this, only the alternative item for Item 20 was added to the final version of the form (Item 14).

The third common criticism was that some items were not appropriate for students’ evaluations of teachers. For example, all teachers and three students criticised Item 7, “Our teacher uses the target language well”, because they noted that students, especially in low proficiency levels, did not have enough knowledge of English to evaluate their teachers’ performance. In response to these comments, Item 7 was omitted on the final version of the form.

A second example for this type of criticism was Item 9, “Our teacher does not bring her personal problems to the classroom”. Although three teachers, who were in favour of the presence of the item on the form, noted that teachers should separate their private life and profession, three teachers suggested that students were not able to identify whether teachers shared their private life inappropriately or whether they used their experiences in their life to make lesson subjects more clear while instructing. For example, one of the teachers noted that she used her noisy neighbour story while instructing the use of “should” while giving advice. In addition to teachers’ criticisms, six students mentioned that reflecting personal problems were normal so they may be tolerated if they did not occur all the time. In response to these comments, Item 9 was eliminated on the final version of the form.

Item 12, “Our teacher makes lessons interesting by considering our interests.” was another example of an item that was not appropriate for students’ evaluations of teachers. Nine students and four teachers found it impossible to find out each student’s interest areas and make lessons interesting in this way in crowded classes. So, they noted that it was unfair to include the item on the teachers’ evaluation. With this in mind, Item 12 was omitted on the final version of the form.

The fourth commonly cited criticism was that some items should be combined instead of using separate items for evaluating similar qualities. They were not considered as overlapping items, yet they were seen as relevant items that can be expressed together. For example, it was suggested that Item 5, “Our teacher gives explanatory examples when necessary” and Item 6, “Our teacher does a sufficient number of exercises and repetition to help us to learn better.” be combined because all these acts were done for reinforcement. With this in mind, a new item “Our

teacher gives explanatory examples, or does a sufficient number of exercises or repeats things to help us to learn better when necessary” (Item 4) was added to the final version of the form.

Another example of this type of criticism was Item 25, “Our teacher gives clear feedback on homework and projects etc.”, Item 26, “Our teacher explains our mistakes on exams in addition to our grades”, and Item 27, “Our teacher explains the correct answers of the exercises we have done”, which were all about evaluating the quality of giving clear feedback., were suggested to be combined by four teachers and four students. In response to these criticisms, a new item “Our teacher clearly explains our mistakes (in homework, in-class exercises, presentations, projects, and the exams that he evaluated etc.)” (Item 19) was added to the final version of the form.

The last commonly cited criticism was some problems in language use or word choice. These problems did not cause ambiguity, but it was felt that they could be expressed better in different words. For example, in Item 28, “Our teacher makes constructive criticisms to improve our language learning process”, the phrase “constructive” was used. During the interviews, five students and five teachers noted that teachers should make constructive criticisms. However five students mentioned that if they received constructive criticism all the time, they started to believe that they knew more than they needed and they stopped studying. Therefore, they suggested that it was better if they received realistic criticisms in appropriate language. In response to these criticisms, a new item “Our teacher uses appropriate and constructive language while commenting on our performance” (Item 20) was added to the final version of the form.

As it was mentioned in the previous section, learner autonomy and liberal culture were also discussed during the interviews with the participants. In terms of learner autonomy, the researcher had to explain the meaning of learner autonomy to the students in classes, and asked whether their teacher provided learner autonomy in their classes. Only three students answered this question and all of them mentioned that there was not learner autonomy in lessons, and teachers took all decisions for them. Moreover, the teachers mentioned that the present education system made it impossible to bring learner autonomy in lessons, and the definition of learner autonomy was not clear. With this in mind, the final form did not include an item for learner autonomy in EFL classes. Moreover, while discussing whether teachers should have liberal culture, all participants suggested that having knowledge of culture should be seen as a whole, so Item 6 in the final form was found enough to measure the cultural knowledge of teachers by considering its relevance to EFL learning.

As a result of the analysis of the interviews about the third version of the form, the items of the final version of the form were designed, 21 items in total (see Figure 6 on page 69). In the following section, the quantitative analysis of the piloting of the final version of the form will be described.

Figure 6 - The final version of the form

1. Our teacher's instruction is clear.
2. Our teacher clearly answers our questions related with the subject.
3. Our teacher adjusts her English by considering our proficiency levels in her lesson.
4. Our teacher gives explanatory examples, or does a sufficient number of exercises or revises subjects to help us to learn better when necessary.
5. Our teacher uses her body language and voice appropriately to strengthen her instructions.
6. Our teacher is able to explain cultural features of the target language when necessary.
7. Our teacher motivates us towards the lesson positively.
8. Our teacher behaves patiently during the language teaching process.
9. Our teacher pays each of us as much attention as we need.
10. Our teacher comes to lessons well prepared.
11. Our teacher uses lesson time efficiently.
12. Our teacher clearly explains why we learn subjects.
13. Our teacher effectively uses group and pair work in lessons.
14. Our teacher effectively handles problems that occur in the classroom.
15. Our teacher uses the course book, board and technological features of the classroom effectively.
16. Our teacher encourages us to participate in lessons.
17. Our teacher explains how useful what we have learned is in real life.
18. Our teacher guides us to do self study outside of the classroom.
19. Our teacher clearly explains our mistakes (in homework, in-class activities, presentation and projects, exams that he/she grade) to improve our language.
20. Our teacher uses appropriate and constructive language while commenting on our performance.
21. Our teacher appreciates our effort and work.

The analysis of the piloting of the final version

This section presents information about the quantitative analysis of the piloting of the final version of the form at AUSFL. The information about participants, distribution of data, reliability of the form, factor analysis, and a comparison of items in different levels and skills will be described in turn.

Participants

The final version of the form was administered to 382 students at AUSFL in 34 different classes, each taught by a different instructor. Nine of the participants missed one or more questions and were deleted from analysis, so 373 participants' forms were analyzed quantitatively using SPSS. Two hundred and sixteen female and 157 male students, whose age distribution is shown in Table 3, participated in the study. Table 4 presents the participants' distribution across the five proficiency levels and four language skills of the teachers that were evaluated.

Table 3 - The age of the student participants in the piloting of the form

Students' age	17-19	20-21	22+
Number of students	210	140	23

Table 4 - Participants' distribution across proficiency level and language skill

Student Proficiency Level	Skill of Teacher being evaluated				Total
	Reading Number of Students/ Number of Classes	Grammar Number of Students/ Number of Classes	Writing Number of Students/ Number of Classes	Speaking & Listening Number of Students/ Number of Classes	
Elementary	26/3	19/1	26/2	35/3	106/9
Low intermediate	34/2	23/2	30/2	35/2	122/8
Intermediate	25/2	38/4	39/3	33/4	135/13
Upper Intermediate	2/1	4/1	0	1/1	7/3
Advanced	0	0	0	3/1	3/1
Total	87/8	84/8	95/7	107/11	373/34

Distribution of the data

The first issue that will be considered is the distribution of responses of the students for each item; that is, whether a full range of responses (from one to five) is given for each item and whether these responses are normally distributed. The importance of this is that it shows whether items play a role in distinguishing teachers' performance or not. If an item does not elicit a range of responses, or if there is skewedness in responses, such that most answers fall at one end of the range, that item may not be useful in distinguishing good teachers from bad teachers.

All items elicited a complete spread of ratings, from one (strongly agree) to five (strongly disagree). However, 16 of 21 items also showed a positive skew, showing that students had a tendency to give positive evaluations about their teachers. Table 5 presents the distribution of the average scores for each item for all 34 teachers. Degree of skewness can be evaluated as Skewness/Standard Error of Skewness, with values of greater than 1.96 deemed to show a significant skew from normality (Field, 2005).

Table 5 - The distribution of responses of the students for each item

Item	Skewness	SE of Skew	Skew/SE	Item	Skewness	SE of Skew	Skew/SE
1	1,47	0,40	3,64	11	0,97	0,40	2,41
2	1,36	0,40	3,38	12	1,48	0,40	3,68
3	1,32	0,40	3,28	13	0,26	0,40	0,63
4	1,66	0,40	4,11	14	1,08	0,40	2,68
5	0,61	0,40	1,51	15	1,15	0,40	2,86
6	0,38	0,40	0,93	16	1,01	0,40	2,51
7	0,95	0,40	2,36	17	0,48	0,40	1,18
8	1,12	0,40	2,78	18	0,44	0,40	1,09
9	1,05	0,40	2,60	19	1,25	0,40	3,11
10	2,31	0,40	5,74	20	1,37	0,40	3,41
				21	1,48	0,40	3,68

As is shown in Figure 7, there are only five items, 5, 6, 13, 17 and 18, shown in bold in the table, which were not significantly skewed. It is worth considering why

these five items achieved a non-skewed distribution (Figure 7). There are several possible explanations for this, and these will be considered in detail in Chapter 5.

Figure 7 - The five normally distributed items

Item 5, Our teacher uses her body language and voice appropriately to strengthen her instructions.
Item 6, Our teacher is able to explain cultural features of the target language when necessary.
Item 13, Our teacher uses group and pair work effectively in lessons.
Item 17, Our teacher explains how we can use what we have learned in real life”.
Item 18, Our teacher guides us to do self-study outside of the classroom.

Reliability

As Table 6 shows, the overall reliability of the form is very high. It is also high when each skill and three levels are considered separately². This shows that the questionnaire elicits information about a coherent construct.

Table 6 - The reliability scores of the final form

	Overall	Skills			Levels			
		Reading	Grammar	Writing	Speaking/ Listening	Elem Pre- int	Int	
Cronbach's Alpha	.96	.97	.96	.94	.94	.98	.95	.94

Factor Analysis

To examine further the nature of the construct evaluated by the questionnaire, a factor analysis was conducted. A Principal Components was used. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (= .972) and the Bartlett's Test of Significance (< .001) indicated that the data met the assumptions of the analysis. The Principal Component Analysis identified only a single factor with an Eigenvalue of greater than 1. This means that students give similar ratings for all questions. This situation may indicate that the construct evaluated is a unitary one. That is to say,

² Upper-Intermediate and Advanced level students' ratings are not shown on the table because of the inadequate number of participation from these levels. There were only 12 students from these levels; and two of them missed one or more items; therefore, only 10 students' ratings were not found adequate enough to be taken into consideration compared to the number of participants from other levels.

good teaching, as it is measured by this form, does not have distinct sub-categories (such as communication skills, field knowledge) about which distinct evaluations can be made.

A possible explanation for this may be that of the halo effect, which means the transfer of feelings or perceptions about one thing to another thing, in students' ratings. In this context, the halo effect occurs when students consider their teacher good (or bad) in one quality; they are likely to make similar evaluation in other qualities. As a result, the form does not provide enough information to draw conclusions about any particular aspects of good teaching.

The differences of the average scores in terms of language skills

A further issue that should be considered is that of differences in the evaluation of teachers in different skill areas. It is possible that some items will be more likely to attract high ratings in some skill classes than in others. For example, it seems likely that an item such as Item 16, "Our teacher gives us the courage to participate in lessons" will typically receive higher scores in Speaking and Listening classes than in Grammar classes.

In the piloting of the form, 34 teachers of four different skills were evaluated in terms of their performances in lessons by 382 students. Each teacher was evaluated once and each student evaluated one teacher. Students rated the items from one (strongly agree) to five (strongly disagree). There were eight classes (87 students) in Reading, eight classes (84 students) in Grammar, seven classes (95 students) in Writing, and 11 classes (107 students) in Speaking and Listening who evaluated their teachers. In order to see whether there were any differences between the average scores of teachers in terms of skills, a one-way independent samples

ANOVA was done. It was found that 16 of 21 items did not favor a specific skill. However, the average scores of five items (2, 5, 13, 16 and 19) were significantly different in terms of skills (see Table 7).

Table 7 - The average scores of teachers in terms of skills

Items	Reading	Grammar	Writing	List. & Speaking	ANOVA
1	1.92	2.72	1.91	1.63	$F(3,33)=2.58$ $p=.072$
2	1.79	2.63	1.72	1.60	$F(3,33)=3.45$ $p=.029$
3	1.90	2.33	1.83	1.57	$F(3,33)=2.07$ $p=.125$
4	2.03	2.34	1.88	1.67	$F(3,33)=1.08$ $p=.371$
5	2.21	2.53	2.05	1.53	$F(3,33)=2.94$ $p=.049$
6	2.33	2.43	2.48	2.06	$F(3,33)=.91$ $p=.447$
7	2.31	3.02	2.30	2.08	$F(3,33)=2.14$ $p=.116$
8	2.02	2.19	1.95	1.65	$F(3,33)=.79$ $p=.504$
9	2.07	2.65	2.23	1.91	$F(3,33)=1.48$ $p=.240$
10	1.80	2.10	1.70	1.54	$F(3,33)=1.09$ $p=.367$
11	2.05	2.46	2.05	1.98	$F(3,33)=.66$ $p=.581$
12	2.15	2.66	2.10	1.94	$F(3,33)=1.60$ $p=.209$
13	2.67	3.55	3.20	1.96	$F(3,33)=9.38$ $p=.000$
14	2.49	2.53	2.42	2.14	$F(3,33)=.53$ $p=.663$
15	2.01	2.33	1.80	1.50	$F(3,33)=2.26$ $p=.102$
16	2.48	2.97	2.23	1.79	$F(3,33)=3.47$ $p=.028$
17	2.77	2.98	2.62	2.33	$F(3,33)=1.42$ $p=.256$
18	2.91	3.23	2.65	2.21	$F(3,33)=2.87$ $p=.053$
19	2.41	2.86	1.93	1.90	$F(3,33)=3.05$ $p=.043$
20	2.29	2.68	2.05	1.88	$F(3,33)=1.49$ $p=.236$
21	2.18	2.46	1.98	1.85	$F(3,33)=1.05$ $p=.384$
OVERALL	61.86	70.41	60.43	54.53	$F(3,33)=2.14$ $p=.115$

The items showing significant differences between skills areas were:

- Item 2: Our teacher answers our questions related with the subject clearly.
- Item 5: Our teacher uses her body language and voice appropriately to strengthen her instructions.
- Item 13: Our teacher uses group and pair work effectively in lessons.
- Item 16: Our teacher gives us the courage to participate in lessons.
- Item 19: Our teacher explains our mistakes (in homework, in-class activities, presentation and projects, exams that he/she grade) clearly to improve our language.

As is shown in Table 7, the teachers of Speaking and Listening have the highest grades and the teachers of Grammar have the lowest ratings in each of these five items. At least three possible interpretations may be made here.

One relates to comments made by teachers during the interviews that, because of the loaded syllabus of the grammar lessons, they had no time to show individual interest for students' questions or mistakes or use extra activities to reinforce the subjects of the course. It is possible that the findings regarding differences between the ratings of different skills reflect this point. Another possible interpretation is that teachers are simply not using appropriate teaching approaches in grammar lessons. This may indicate a need for training to improve their instruction. Finally, it may be that these items naturally favor the types of teaching found in certain skill classes. This would suggest that the form should ideally not be used to evaluate all skills and that adapted versions of the form specially suited to each skill should be developed.

Conclusion

The chapter presents the findings of both the qualitative and quantitative analysis of the present study. The next chapter will present a discussion of the research findings concerning the research questions and the relevant literature, the limitations of the study, the pedagogical implications derived from the results and suggestions for further studies.

CHAPTER V- CONCLUSION

Introduction

The purpose of this study was to design a valid EFL- specific TEF for student ratings which evaluates the performance of language teachers. The researcher has attempted to answer the following questions:

1. What items should be included in an EFL-specific teacher evaluation form for student ratings?
 - a. To what extent is the evaluation form currently in use at AUSFL satisfactory?
 - b. To what extent are concepts from the literature on effective teacher behaviors a good basis for creating a teacher evaluation form?
 - c. Is the construct of ‘good teaching’, as evaluated by students, a unitary construct or can it be divided into distinct sub-categories?
 - d. Are there differences in the ways in which students evaluate teachers of different language skills (such that evaluation forms ought to be made skill-specific)?

This study was conducted with 392 students and 21 teachers at AUSFL and 3 teachers from different universities. An initial version of the evaluation form was created based on the data gathered from literature review and an online questionnaire with 15 teachers. (Six of these 15 teachers also participated in the further interviews on the sample form.) Then four teachers were interviewed on the initial version of the form and the second version of the form was designed. Five teachers were interviewed on the second version of the form, and the data gathered from these interviews was used to design the third version of the form. Six teachers and ten

students were interviewed on the third version of the form, and the data gathered from the third set of interviews was used to create the final version of the form. The three sets of interviews about the form and the answers of the online questionnaire were analyzed qualitatively by the researcher. The final version of the form was piloted with 382 students in 34 classes from different levels and skills and analyzed quantitatively. The researcher paid attention to balance the number of students in terms of skills and levels. Nine students missed one or more questions on the form; therefore, these papers were eliminated in the quantitative analysis. The whole procedure was conducted in Turkish.

This chapter presents the research findings concerning the research questions and the relevant literature, the limitations of the study, the pedagogical implications derived from the results and suggestions for further studies.

General results and discussion

The final version of the form is the answer to the question of what items should be included in an EFL-specific TEF for students' ratings. We can get an idea of the usefulness of this form by looking at the designing process of the form, the reliability, and the distribution of the ratings. Regarding the first, the method of designing the form should have ensured its appropriateness for measuring language teaching. The initial form was based on the effective teacher behavior literature and the distinctive features of language teaching. Both literatures provided information about the needs of the EFL classroom. The participants in the interviews, who were all EFL teachers and students at AUSFL, provided information about practice in the classroom and the needs of language teaching from different aspects. All items on the form were discussed in detail during the interviews, and the items that were

accepted by majority of the participants were put on the form. Therefore, the process itself was focused on the needs of language teaching. Second, regarding reliability, the overall reliability of the form ($\alpha=0.96$) is very high, indicating that the form measures a coherent construct. Thirdly, and less encouragingly, the distributions of the ratings of 16 items were positively skewed, which could be explained in many ways. The skewedness of these items may not distinguish good teachers from bad teachers. If students gave positive ratings for all teachers, it means that the form could not provide any specific information about teaching performance with these items. Another possible explanation for the skewedness was that the teachers who were willing to be rated by their students were teachers who were good at their jobs, so it is natural to have positively skewed results. Moreover, the skewedness may be a result of a halo effect in students' ratings. Students may have scored the items according to their first impression, instead of paying attention to what each item measured specifically. Therefore, if students had a positive attitude towards their teacher, they gave high scores to each item. On the other hand, there were five items which were normally distributed. It may be inferred that these five items were able to distinguish good teachers from bad teachers. Another explanation for these findings was that these may have been related to more observable behaviors from students' perspectives, so students were able to identify whether their teachers did what these normally distributed items measure. Last, these items may reflect students' expectations from language learning, so they had a clearer idea about what teachers should do in terms of these normally distributed items. These findings could be the limitation of the form in this study. With this in mind, the form can be considered as a good starting point for creating an EFL-specific TEF for students' ratings.

Regarding research question (a), the current TEF for students' ratings at AUSFL was not found satisfactory according to analysis of the fourth question of the online questionnaire administered at the beginning of the study. Twelve teachers from AUSFL made two basic types of criticism of the current form. First, they stated that it included items that were not under teachers' control, so it was not fair to judge them with these items. Burden (2008) also criticized some items in TEFs, which measured various aspects of an EFL classroom, such as the syllabus, students' self-evaluation, the course book and supplementary material, and classroom equipment, for being not related with teachers' ability to teach. The second criticism was the ambiguity of certain items which may lead students to misinterpret them. The participants suggested that items on the form should specify what they measure clearly.

Another discussion of the present form could be regarding the consequential validity of the findings. The results of the present students' ratings could be seen on the system by administrators who mentioned that they checked the results each term, yet there has not been any clear explanation about the effects of the findings at the institution. Teachers can only see their own results if at least twelve students rate their performance in a classroom. Administrators have the right to see the results of all teachers in their department; however in practice, the results do not have any formative or summative outcomes. Moreover, the present research found that teachers did not believe the present form adequate to provide information about teachers' performance. It is hoped that this form will be a step towards providing a more informative and clear form to the institution in order to use the form more effectively. The items of the new form were chosen from the literature of effective

teacher behaviours and the distinctive features of EFL teachers, and discussed and shaped by EFL students and teachers during the study; therefore, the results of this evaluation form can be considered more reliable by teachers who are willing to improve their teaching performance; and by administrators who may use the information gathered from the evaluation form to organize internal training for their teachers. However, although it may be possible to use the information gathered from the evaluation form summatively, this may not be the best policy. In the long run this situation may be abused by students who may threaten their teachers with giving low grades and it may affect teachers' performance and the quality of education negatively because teachers may have to curry favour with students in order not to lose their jobs. The best use for this form may rather be formative. Only in this way fair judgment of students could be possible and teachers may be more open to accepting the positive and negative criticism of students' ratings; and improve their teaching performance.

Regarding research question (b), it was found that concepts from the effective teacher behaviors literature provide a good basis to design a TEF for students' ratings. However, it is important to note that, while the ETB literature provides a good source of ideas, it does not by itself provide items necessary to creating a form. At least two points need to be considered when basing evaluation forms on this literature. First, though the ETB literature gives a general idea of 'effective teaching' in all fields, as Neuman (2001) mentioned, the disciplines may differ from one another in terms of several factors, such as types of teaching, preparation time, practice, curriculum, assessment of students, program review, teaching approaches, and teaching outcomes. With this in mind, the ETB literature was not sufficient to

create a TEF for language teachers. Burden (2008) also suggested that language teachers should be evaluated by an EFL-specific TEF. For this reason, the literature regarding the distinctive qualities of language teachers was taken into account while designing the form. The distinctive features of language teaching provide a good deal of information for choosing appropriate qualities from the ETB literature. Second, another important aspect of creating a new form is to design items which measure expected quality. Dörnyei (2003) suggested that wording plays an important role while designing a form. With this in mind, to transform ideas from the literature into a form, the researcher studied how to write a questionnaire appropriately. During the interviews in the study, one of the primary concerns of the researcher was to find appropriate wording for measuring the target quality. For this reason, items were discussed throughout the interviews as to whether they were ambiguous, overlapping, or not appropriate for students' evaluation of teachers, or whether there were other wording and language use problems. Each set of interviews provided information about items that should be modified and clarified.

Regarding research question (c), the factor analysis of the form in this study showed that the construct of 'good teaching', as evaluated by students, appears to be a unitary construct. It is possible to explain this result in two ways, whether the form in this study may not be adequate to distinguish the multidimensional aspects of language teaching, or the findings of the present study may show that students had a tendency to give similar ratings to every item. This means that instead of rating each item separately, students may have rated their teachers according to their general impressions. This is an unexpected result because several studies in the literature suggested that students' ratings are multidimensional (Marsh & Roche, d'Apollonia

& Abrami and Greenwald & Gillmore, as cited in Greenwald, 1997). This may be interpreted as a halo effect, which means that the impressions of successful performance in one area are translated to overrating in another (Grussing, 1994). This would suggest that students do not rate teachers according to the specific questions asked, but rather on the basis of an overall impression. These findings may indicate that students' ratings are only a measure of teacher popularity (Nasser & Fresko, 2002). If this is the case, it may not be possible to collect detailed information about different aspects of a teacher's performance; this would also suggest that long and detailed evaluation forms are unnecessary. On the other hand, it is possible that a differently-designed form may be able to elicit separate information about different aspects of good teaching. This is an area that warrants further research.

Finally, for research question (d), the study also explored whether there are differences in the ways in which students evaluate teachers of different language skills. The analysis of students' responses to the form showed that for five items there were significant differences in the scores given to teachers from different skill areas. In particular, listening and speaking teachers were rated particularly high and grammar teachers particularly low on these items. It is very difficult to find certain explanations for the reasons of this situation. One possible conclusion from this situation is that teachers mentioned that in grammar lessons the syllabus is very loaded; therefore, they do not have time for answering students' questions, correcting their mistakes, using group and pair work, and reinforcing the subjects. Another possible conclusion is that teachers may not teach by considering effective teaching behaviors in grammar lessons. Therefore, the institution may provide in-service

training on skill based teaching to improve teachers' instruction. These findings could also indicate that the form is not adequate to evaluate teachers' performances in grammar. For this reason skill-based TEFs may be needed to evaluate teaching performance efficiently. This takes the forms one step beyond Burden's (2008) suggestion about an EFL specific TEF for students' ratings. Further studies should be conducted on whether each language skill needs a specific form for teacher evaluation for students' ratings.

Another discussion raised at the end of the study was the extent to which the final version of the form was EFL specific. At the end of the study, the form consisted of 21 items for evaluating language teachers' performance by students' ratings. The items on the form were derived from the literatures on effective teacher behaviors and the distinctive features of EFL teachers, an online questionnaire and semi-structured interviews with students and teachers which all provide information about the needs of EFL classroom in terms of teaching practice. Although language teaching has some distinctive features, it is expected to have some overlaps with other disciplines in terms of good teaching performance. Being a good language teacher is a subcategory of being a good teacher. Therefore, the form considers the needs of language teaching with all aspects of good teaching in general. For example, every teacher should provide clear instruction in any disciplines, so language teachers should provide clear instruction, too. Therefore, Item 1 on the final version of the form measures the clarity of the instruction of the form.

Implications

The first implication of the study is that the current online TEF of AUSFL is not satisfactory; therefore, the main practical outcome of the study is the TEF itself. The existing form should, I believe, be replaced. The form developed in this study is offered as a draft replacement.

The second implication is that the literature on ETB provides a good basis to design teacher evaluation forms. However, while designing a new form for a specific field, the distinctive features of the target field should be considered. Semi-structured interviews with stakeholders also provide a great deal of data to identify items that are overlapping, ambiguous, not appropriate for students' evaluation of teachers and subjects that are neglected. These points should be taken into consideration in any future form development projects.

The third implication is that 'good teaching' appears to be a unitary construct. As discussed above, one possible practical outcome of this situation is that we may not need to use sub-categories to measure teachers' performance. This means that a few items may be enough to measure students' impression about teaching performance if they give similar ratings to all items. Moreover, if the form only provides students' general impressions about teaching performance, it is important that these forms not be considered as a final judgment on teachers' performance, though they should be considered as a part of evaluation of teachers' performance.

Lastly, the findings showed that the teachers of grammar at AUSFL scored in some items significantly lower compared to other skills. One possible explanation for this situation is the lack of time because of the loaded syllabus in grammar lessons, as teachers mentioned during the interviews. If this is the explanation, the

grammar syllabus should be revised to provide more time for teachers. Another possible explanation is that teachers may not teach by considering effective teaching behaviors in grammar lessons. Therefore, the institution may provide in-service training on skill-based teaching to improve teachers' instruction. This situation also may show the need for skill-specific forms to evaluate teachers' performance in different skills.

Limitations of the study

One of the limitations of this study is the choice of the participants. As Theall and Franklin (as cited in Onwuegbuzie et al., 2007) mentioned, all stakeholders should be included in the evaluation process, so that the perspectives of all stakeholders can be considered (Ory and Ryan, as cited in Onwuegbuzie et al., 2007). However, in the present study only students' and teachers' perspectives about the items on the TEF were considered. The point of views of the administrators should also be taken into account while designing a TEF. Moreover, the spread of proficiency levels of students who participated in the study could not be done in a balanced way; many high proficiency level students had passed the proficiency exam at the end of the first term, yet the interviews with students were held at the end of the second term. For this reason, the number of the students from upper-intermediate and advanced level is only 10, so the findings of the study were mainly based on the elementary, lower-intermediate and intermediate level students' and teachers' views.

Another limitation of the study is that the distributions of the ratings of 16 items were positively skewed, which means that these items were not able to distinguish good teachers from bad teachers. Only five items were normally distributed. It may be inferred that only these five items were able to distinguish

good teachers from bad teachers. The reasons for this situation could have been identified with some interviews after the piloting.

Suggestions for further research

Bearing in mind the limitations of the study, a number of suggestions could be made for further studies. A similar study could be conducted which includes the views of administrators, so that the research may be constructed on all stakeholders' opinions. The number of students from different proficiency levels should be balanced to gather different views of students from different levels equally. Including all stakeholders' views and balancing the participation of students from different proficiency levels should be considered in further studies.

Moreover, the findings of the study show that there may be a need for skill-specific forms in language teaching. To what extent skills differ in terms of teachers' performance should be explored in further studies.

Lastly, the factor analysis of the form shows that 'good teaching' may be a unitary construct. This situation may be the result of the inadequacy of the form in this study to evaluate multidimensional aspects of language teaching or it may be evidence for the halo effect in students' ratings. The reasons for this finding should be explored in further studies.

Conclusion

The present study aimed to design a valid EFL-specific TEF for students' ratings. Ten students and 24 teachers were interviewed in the study, and the final version of the form was piloted at AUSFL with 382 students from different levels and skills. The study provided a basis for an EFL-specific TEF for students' ratings. The findings of the research suggested that the current TEF for students' ratings was

not sufficient; the ETB literature provided a good basis for designing a new TEF, yet the distinctive features of language teaching should be also considered while designing an EFL-specific TEF; according to the factor analysis of the form in this study, the construct of ‘good teaching’ emerged as a unitary construct, but this may be due to the inadequacy of the form in this study to measure multiple dimensions of language teaching, or the presence of a halo effect in students’ ratings; and the scores of teachers were significantly different when they were compared with each other in terms of skills. In conclusion, the final form that was designed in the study could be considered as a good draft for creating an EFL-specific TEF for students’ ratings.

REFERENCES

- Adams, J. (1997). Student evaluations: The ratings game. *Inquiry*. Retrieved from <http://www.vccaedu.org/inquiry/inquiry-fall97/i12-adam.html>.
- Alhija, F. N., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, 35(1), 37-44.
- Bell, T. R. (2005). Behaviors and attitudes of effective foreign language teachers: questionnaire study. *Foreign Language Annals*, 38(2), 259-273.
- Block, D. (1998). Exploring interpretations of questionnaire items. *System*, 26(3), 403-425.
- Brown H. D. (2004). *Language assessment: principles of language assessment*. NY: Longman.
- Borg, S. (2006). The distinctive characteristics of foreign language teachers. *Language Teaching Research*, 10(1), 3.
- Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9), 1-5.
- Burden, P. (2008). ELT teacher views on the appropriateness for teacher development of end of semester student evaluation of teaching in a Japanese context. *System*, 36(3), 478-491.
- Center of Enhanced Learning and Teaching (2010). Student feedback questionnaires. Retrieved December 12, 2010, from <http://celt.ust.hk/instr/instr11.htm>.
- Çakmak, M. (2009). Öğretmen Adaylarının “ Etkili Öğretmen ” Nitelikleri Konusunda Düşünceleri [Prospective Teachers ' Thoughts on Characteristics of an “ Effective Teacher ”] . *Eğitim ve Bilim [Education and Science]*, 34(153).
- Devlin, M., & Samarawickrema, G. (2010). The criteria of effective teaching in a changing higher education context. *Higher education research and development*, 29(2), 14. Routledge.
- Dondanville, R. A. (2005). *Assessing effective teacher behaviors in athletic training clinical education*. The University of North Carolina at Greensboro. Retrieved from <http://libres.uncg.edu/ir/uncg/listing.aspx?styp=ti&id=891>.
- Dörnyei Z. (2003). *Questionnaires in second language research: Construction, administration, processing*. New Jersey: Lawrence Erlbaum Associates Publishers.

- Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1), 37-46.
- Field, A. (2005). *Discovering statistics using SPSS (Second ed.)*. London: Sage Publications.
- Greenwald, A. (1997). Validity concerns and usefulness of student ratings of instruction. *The American Psychologist*, 52(11), 1182-1186.
- Grussing, P. G. (1994). Sources of error in student evaluation of teaching. *American Journal of Pharmaceutical Education*, 58(3), 316-318.
- Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *Test*, 8(4), 597-606.
- Hughes A. (1998). *Testing for language teachers (First ed.)*. Cambridge: Cambridge University Press.
- Hughes A. (2004). *Testing for language teachers (Second ed.)*. Cambridge: Cambridge University Press.
- Huitt, W. (2003). A transactional model of the teaching/learning process. *Educational Psychology Interactive*. Retrieved from <http://www.edpsycinteractive.org/materials/tchlrmnd.html>
- Jang, S.-J., Guan, S.-Y., & Hsieh, H.-F. (2009). Developing an instrument for assessing college students' perceptions of teachers' pedagogical content knowledge. *Procedia - Social and Behavioral Sciences*, 1(1), 596-606.
- Kyale, S. (1996). *Interviews on introduction to qualitative research interviewing*. California: Sage Publication.
- Kidd, R. (2004). Student evaluations: Are they valid measures of course effectiveness. *American Journal of Pharmaceutical Education*, 68(3), 3-8.
- Lee, J. J. (2010). The uniqueness of EFL teachers: Perceptions of Japanese learners. *TESOL Journal*, 1(1), 23-48.
- Mace, J. P. (1997). Valuing the individual student: Using single-subject design as a tool for evaluating classroom teaching performance. *Journal of Social Work Education*, 33(2), 361-74.
- Madden, T. J., Dillon, W. R., & Leak, R. L. (2010). Students' evaluation of teaching: Concerns of item diagnosticity. *Journal of Marketing Education*, 32(3), 264.

- Marlin J. W. Jr. & Niss J. F. (1980). End-of-course evaluations as indicators of student learning and instructor effectiveness. *The Journal of Economic Education*, 11(2), 16-27.
- Marsh, H. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707-754.
- McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, 52(11), 1218-1225.
- Murdoch University (2004). Student surveys of teaching. Retrieved December 12, 2010, from <http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html>.
- Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, 27(2), 187-198.
- Fresko, B. & N. F. (2001). Interpreting student ratings: Consultation, instructional modification, and attitudes towards course evaluation. *Studies in Educational Evaluation*, 27, 291-305.
- Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, 26(2), 135-146.
- Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation from using a mixed-methods Analysis. *American Educational Research Journal*, 44(1), 113-160.
- Question Pro (n.d.). Student course evaluation –teacher evaluation. Retrieved December 17, 2010, from <http://www.questionpro.com/akira/showSurveyLibrary.do?surveyID=88&mode=1>
- Rahimi, M., & Nabilou, Z. (2010). Iranian EFL teachers' effectiveness of instructional behavior in public and private high schools. *Asia Pacific Education Review*, 12(1), 67-78.
- Riddle D. (n.d.) Ten tips for designing training evaluation forms, with samples. Retrieved November 30, 2010, from <http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf>.
- Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, 39, 40.

- Shishavan, H., & Sadeghi, K. (2009). Characteristics of an effective english language teacher as perceived by Iranian teachers and learners of English. *English Language Teaching*, 2(4), 130-143.
- Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.
- Spencer, K. & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, 27(5), 397-409.
- Sahenk, S. (2010). Characteristics of the headmasters, teachers and students in an effective school. *Procedia-Social and Behavioral Sciences*, 2, 4298-4304.
- Taylor, R., Reeves, B., Mears, R., Keast, J., Binns, S., Ewings, P., & Khan, K. (2001). Development and validation of a questionnaire to evaluate the effectiveness of evidence-based practice teaching. *Medical education*, 35(6), 544-7.
- University of Sussex. (2010). Course evaluation questionnaires item bank. Retrieved December 17, 2010, from <http://www.sussex.ac.uk/tldu/ideas/eval/ceq>.
- University of Ulster (2008). Assessment of teaching - student questionnaire report 2006-2007. Retrieved December, 17, 2010, from <http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire>.
- Wei L. (2010). Formative assessment: Opportunities and challenges. *Journal of Language Teaching and Research*, 1(6), 838-841.
- Wennerstrom, A. K., & Heiser, P. (1992). ESL student bias in instructional evaluation. *TESOL Quarterly*, 26(2), 271-288.
- Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *The Journal of Education for Business*, 78(6), 313-317.

APPENDIX A - FINAL VERSIONS OF THE FORM, ENGLISH AND TURKISH

English Version of the Final Form

1. Our teacher's instruction is clear.
2. Our teacher clearly answers our questions related with the subject.
3. Our teacher adjusts her English by considering our proficiency levels in her lesson.
4. Our teacher gives explanatory examples, or does a sufficient number of exercises or revises subjects to help us to learn better when necessary.
5. Our teacher uses her body language and voice appropriately to strengthen her instructions.
6. Our teacher is able to explain cultural features of the target language when necessary.
7. Our teacher motivates us towards the lesson positively.
8. Our teacher behaves patiently during the language teaching process.
9. Our teacher pays each of us as much attention as we need.
10. Our teacher comes to lessons well prepared.
11. Our teacher uses lesson time efficiently.
12. Our teacher clearly explains why we learn subjects.
13. Our teacher effectively uses group and pair work in lessons.
14. Our teacher effectively handles problems that occur in the classroom.
15. Our teacher uses the course book, board and technological features of the classroom effectively.
16. Our teacher encourages us to participate in lessons.
17. Our teacher explains how useful what we have learned is in real life.
18. Our teacher guides us to do self study outside of the classroom.
19. Our teacher clearly explains our mistakes (in homework, in-class activities, presentation and projects, exams that he/she grade) to improve our language.
20. Our teacher uses appropriate and constructive language while commenting on our performance.
21. Our teacher appreciates our effort and work.

Turkish Version of the Final Form

Öğrenciler İçin İngilizce Öğretmeni Performans Değerlendirme Formu	
Soruları cevap anahtarına kodlayınız.	
Cevaplarınız;	A=Kesinlikle Katılıyorum B=Katılıyorum C=Kısmen Katılıyorum D=Katılmıyorum E= Kesinlikle Katılmıyorum şeklinde değerlendirilmeye alınacaktır.
1.	Öğretmenimizin ders anlatımı açık, net ve anlaşılırdır.
2.	Öğretmenimiz konuyla ilgili sorulara açıklayıcı ve anlaşılır cevaplar verir.
3.	Öğretmenimiz ders anlatırken İngilizceyi seviyemizi göz önüne alarak kullanır.
4.	Öğretmenimiz gerektiği yerlerde konuyla ilgili açıklayıcı örnekler vererek/ alıştırma yaparak/ konuyu tekrar anlatarak konuyu daha iyi anlamamızı sağlar.
5.	Öğretmenimiz ses tonunu ve beden dilini ifadelerini güçlendirecek şekilde kullanır.
6.	Dil öğretim sürecinde gerektiğinde öğrettiği dilin kültürel özelliklerini açıklayabilecek bilgiye sahiptir.
7.	Öğretmenimiz ders motivasyonumuzu olumlu bir şekilde sağlar.
8.	Öğretmenimiz dil öğretim sürecinde bize karşı sabırlı davranır.
9.	Öğretmenimiz derste her birimize ihtiyaç duyduğumuz kadar ilgi gösterir.
10.	Öğretmenimiz derse hazırlıklı gelir.
11.	Öğretmenimiz ders zamanını verimli bir şekilde kullanır.
12.	Öğretmenimiz derste neyi niçin işlediğimizi açık ve net şekilde ortaya koyar.
13.	Öğretmenimiz derste grup ve ikili (pair) çalışmalarını etkili bir şekilde kullanır.
14.	Öğretmenimiz sınıf ortamını bozan sorunlarla etkili bir şekilde baş eder.
15.	Öğretmenimiz ders materyallerini (tahta, ders kitabı, teknolojik donanım vb.) verimli bir şekilde kullanır.
16.	Öğretmenimiz ders katılımımızı sağlamak için bizi cesaretlendirir.
17.	Öğrendiklerimizin gerçek yaşamda ne işimize yarayacağını açık bir şekilde ortaya koyar.
18.	Öğretmenimiz sınıf dışında kendi başımıza çalışabilmemiz için bize yol gösterir.
19.	Öğretmenimiz dil öğreniminde ilerlememiz için (ödevler, sınıf içi alışırmalar, sunum ve projeler, kendi değerlendirdiği sınavlar vb. durumlarda) yaptığımız hataları açık ve anlaşılır şekilde bize açıklar.
20.	Öğretmenimiz dil öğreniminde ilerlememiz için yaptığı eleştirilerde uygun ve yapıcı bir dil kullanır.
21.	Öğretmenimiz çalışmalarımızı ve çabalarımızı takdir eder.
22.	Seviyenizi işaretleyiniz. A) Elementary B) Low-Int C) Int D) Up-Int E)Advanced
23.	Yaş aralığınızı işaretleyiniz. A) 17-19 B) 20-21 C) 22+
24.	Cinsiyetiniz. A) Kız B) Erkek
25.	Değerlendirilmeye aldığımız öğretmenin dersini işaretleyiniz. A) Reading B)Grammar C)Writing D)Speaking/Listening

APPENDIX B - AN EXAMPLE TRANSCRIPT OF THE INTERVIEWS

Türkçe

Görüşmenin başında araştırmacı tarafından yapılan açıklama

Araştırmacı: Bu maddeler üzerinde senden istediğim maddelerden ne anladığın, madde açık ve anlaşılır mı, benzer maddeler dikkatini çekerse var mı, madde formda olmalı yada olmamalı diye yorum yapman.

Görüşme sırasında

Araştırmacı: Öğretmenimiz ses tonunu ve beden dilini etkili biçimde kullanır.

Yüksek düzey İngilizce seviyesindeki öğrenci: Bir hoca için önemli, bazen çok hikâye tarzında anlatıyorlar, uyuyorsun, sonra birden ses yükseliyor uyanıyorsun rüya gibi. İngilizce için bence önemli, diğer öğretmenler için değil de, yabancılar sürekli beden dili ve mimikleri çok kullanıyorlar, bazen o kelimeyi bilemesem bile yaptığı bir hareketle ne dediğini anlıyorum. Özellikle geçen dönem hiçbir şey anlamıyordum çünkü sürekli İngilizce konuşuluyordu ve ben başlangıç seviyesinden gelmiştim ve üç ay boyunca süre geçmişti aradan, bildiğim kelimeleri de unutmuştum ama ondan sonra vücut diliyle İngilizceyi anladım yani.

English

The explanation made by the researcher before the interview

The researcher: Please comment on the items, whether the items are clear, whether there are similar items, or whether the items should be on the form.

During the interview

The researcher: Our teacher uses his intonation and body language efficiently.

Upper intermediate level student: It is important for a teacher, sometimes they instruct as a tale, I sleep, then suddenly the intonation raises, I wake up, it is like a dream. It is important for English, not for other teachers. Foreigners always use their intonation and body language. Sometimes even if I do not know the word, I can understand what he says from his body language. Especially, last term I did not understand anything because they spoke in English all the time and my previous level is elementary and it had been three months so I forgot the words I know. But then I understood English by body language.

APPENDIX C - CURRENT TEF FOR STUDENTS' RATINGS AT AUSFL

1. The objectives of the course being stated clearly.
2. The subjects being taught according to the objectives of the course.
3. The appropriateness of the course book and other materials to the objectives of the course.
4. The appropriateness of homework, projects and other activities to the objectives of the course.
5. The appropriateness of exams to the subjects of the course.
6. The teacher's being prepared for the lesson.
7. The ability of exams to evaluate students' success.
8. The teacher's punctuality.
9. The lessons being conducted in an interesting way.
10. Subjects being taught clearly.
11. The lessons being conducted in a way that fosters participation.
12. The teacher's being open to questions and comments.
13. The teacher's having regular office hours.
14. The teacher's tolerance towards students.

APPENDIX D - PUBLISHED SOURCES FOR EACH QUALITY AND ITEM

Quality in the literature	Reference	Categories/Items
Personal Characteristics	Spencer & Schmelkin, as cited in Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier & Moore (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie, et al. (2007) Marlin & Niss (1980)	Personal Characteristics
Being understanding / empathy	Marlin & Niss (1980) Alhijja & Fresko (2009)	Our teacher behaves in an understanding way towards us.
Caring for students	Okpala & Ellis, as cited in Onwuegbuzie, et al. (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie, et al. (2007) Marsh (1984) Shishavan & Sadeghi (2009) (http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm)	Our teacher is interested in my problems.
Communicator	Spencer & Schmelkin, as cited in Onwuegbuzie, et al. (2007) Kane, Sandretto & Health, as cited in Onwuegbuzie, et al. (2007) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Cruikshank, Bainer and Metcalf, as cited in Şahenk (2010)	Our teacher is able to communicate us efficiently.
Providing a peaceful environment	(http://celt.ust.hk/instr/instr11.htm) Alhijja & Fresko (2009) Jang, Guan & Hsieh (2009) Şahenk (2010) Shishavan & Sadeghi (2009)	Our teacher provides a peaceful environment in the classroom.
Friendly	Marsh (1984) Cruikshank, Bainer and Metcalf, as cited in Şahenk (2010) Shishavan & Sadeghi (2009)	Our teacher behaves like a friend towards us.
Having positive attitude	Cruikshank, Bainer and Metcalf, as cited in Şahenk (2010)	Our teacher has a positive attitude towards us.
Patient	(http://www.edpsycinteractive.org/materials/tchlrnmd.html) Shishavan & Sadeghi (2009)	Our teacher is patient while correcting our mistakes.
Field Knowledge	Kane, Sandretto & Health, as cited in Onwuegbuzie, et al. (2007) Content Knowledge - Okpala & Ellis, as cited in Onwuegbuzie, et al. (2007) Knowledge of Subject- Emery, Kramer & Tian (2003)	Field Knowledge

Having field knowledge	Buskit, as cited in Onwuegbuzie, et al. (2007) Cruikshank, Bainer and Metcalf, as cited in Şahenk (2010) Madden, Dillon & Leak (2010) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Jang, Guan & Hsieh (2009)	Our teacher has enough field knowledge.
Having knowledge in other disciplines	Bell (2005)	Our teacher's knowledge in other disciplines affects language learning positively.
Having knowledge about target languages' culture	Shishavan & Sadeghi (2009) Bell (2005)	Our teacher teaches not only language but also its culture.
Having knowledge about target languages' culture	Bell (2005)	Our teacher gives importance to teaching language with its culture.
Making explanations	(http://celt.ust.hk/instr/instr11.htm).	Our teacher is able to make necessary explanations while instructing.
Answering students' questions	(http://www.questionpro.com/akira/showSurveyLibrary.do?surveyID=88&mode=1)	Our teacher has the knowledge to answer our questions about the lesson.
Improving himself in his profession	Cruikshank, Bainer and Metcalf, as cited in Şahenk (2010) Shishavan & Sadeghi (2009)	Our teacher improves herself/himself all the time in her/his field.
Giving examples	Greilmen-Furhman & Geyer as cited in Onwuegbuzie, et al. (2007) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Jang, Guan & Hsieh (2009) (http://www.edpsycinteractive.org/materials/tchlrnmd.html)	Our teacher is able to give necessary examples which are relevant to the lesson.
The use of L1+L2	Bell (2005)	Our teacher can explain the differences and similarities between Turkish and English Language.
Pedagogical Skills	Kane, Sandretto & Health as cited in Onwuegbuzie, et al. (2007) Presentation skills- Crumbley, Henry, and Kratchman as cited in Onwuegbuzie, et al. (2007) Teaching Skills- Okpala & Ellis as cited in Onwuegbuzie, et al. (2007)	Pedagogical Skills

Being prepared	Crombley, Henry & Kratchman as cited in Onwuegbuzie, et al. (2007) Emery, Kramer & Tian (2003) (http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm) Madden, Dillon & Leak (2010) McGrath, Yeung, Comfort & McMillan (2005) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Marsh (1984)	Our teacher comes to class prepared.
Stating objectives	(http://www.sussex.ac.uk/tldu/ideas/eval/ceq) McGrath, Yeung, Comfort & McMillan (2005) Marsh (1984) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Jang, Guan & Hsieh (2009) Madden, Dillon & Leak (2010) Shishavan & Sadeghi (2009)	Our teacher presents the objectives of the lesson clearly. Since the objectives of the lesson are clear, my learning process becomes meaningful.
Passing on his knowledge to students	Was added to the form based the information gathered from online form	Our teacher is able to pass on her/his knowledge to us.
Answering questions	(http://www.questionpro.com/akira/showSurveyLibrary.do?surveyID=88&mode=1) Wennerstorm & Heiser (1992) Jang, Guan & Hsieh (2009) (http://www.edpsycinteractive.org/materials/tchlrmd.html)	Our teacher answers my questions about the lesson.
Having clear explanations	Greilmen-Furhman & Geyer, as cited in Onwuegbuzie, et al. (2007) Marsh (1984) Wennerstorm & Heiser (1992) Alhijja & Fresko (2009) (http://www.edpsycinteractive.org/materials/tchlrmd.html)	Our teachers' explanations/instructions are clear.
Using course material appropriately	Use the course material in an interesting way	Our teacher uses the course materials well.
Teaching in an interesting way	Buskit, as cited in Onwuegbuzie, et al. (2007) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Wennerstorm & Heiser (1992) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf) Alhijja & Fresko (2009) Jang, Guan & Hsieh (2009) Bell (2005)	Our teacher teaches in an interesting way.

Motivating Students	McGrath , Yeung, Comfort & McMillan (2005) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) Shishavan & Sadeghi (2009)	Our teacher motivates us to learn English.
Encouraging participating	(http://www.sussex.ac.uk/tldu/ideas/eval/ceq) (http://celt.ust.hk/instr/instr11.htm) McGrath , Yeung, Comfort & McMillan (2005) (http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire) Alhija & Fresko (2009) Wennerstorm & Heiser (1992) (http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf)	Our teacher encourages us to participate in the activities.
Teaching studying independently out of classroom	McGrath , Yeung, Comfort & McMillan (2005) (http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html) Bell (2005)	Our teacher teaches us to study independently outside of the classroom.
Providing sufficient practice	Observable behaviour for being prepared	Our teacher gives us sufficient practice.
Making real-life connections	Kane, Sandretto &Health, as cited in v McGrath , Yeung, Comfort & McMillan (2005) (http://www.edpsycinteractive.org/materials/tchlrnmd.html) Bell (2005)	Our teacher uses activities that are like real-life situations.
Using relevant activities	Bell (2005)	Our teacher uses activities that encourage using what we have learned.
Considering students proficiency levels while instructing	Shishavan & Sadeghi (2009) Bell (2005)	Our teacher explains the lesson in a way that is suitable to our proficiency level.
Guiding students to study out of classroom	McGrath , Yeung, Comfort & McMillan (2005) Shishavan & Sadeghi (2009) Bell (2005)	Our teacher guides us to study out of classroom.
Feedback &Assessment	A part of pedagogical skills category, yet the researcher decided to add a new category for evaluating feedback and assessment qualities.	Feedback & Assessment
Assessing what has been taught	<u>Shishavan & Sadeghi (2009)</u>	Our teacher assesses what we have learned in the lesson by the quizzes.

Giving explanatory /useful feedback	http://www.sussex.ac.uk/tldu/ideas/eval/ceq McGrath , Yeung, Comfort & McMillan (2005) http://www.tlc.murdoch.edu.au/eddev/evaluation/survey/teachdraft.html http://www.edpsycinteractive.org/materials/tchlrnmd.html Bell (2005)	Our teacher gives explanatory feedback to our homework.
Giving constructive feedback	http://www.edpsycinteractive.org/materials/tchlrnmd.html	Our teacher corrects our mistakes without making us feel bad.
Giving feedback in time	http://celt.ust.hk/instr/instr11.htm http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf Madden, Dillon & Leak (2010) http://www.edpsycinteractive.org/materials/tchlrnmd.html	Our teacher gives feedback on time.
Being fair while assessing	Crumbley, Henry, and Kratchman, as cited in Onwuegbuzie, et al. (2007) Buskit (2003) Emery, Kramer & Tian (2003) http://celt.ust.hk/instr/instr11.htm http://www.edpsycinteractive.org/materials/tchlrnmd.html http://staffdev.ulster.ac.uk/index.php?page=assessment-of-teaching-student-questionnaire http://www.servicegrowth.net/documents/10%20Tips%20on%20Creating%20Training%20Evaluation%20Forms.net.pdf	Our teacher is fair while evaluating our in-class performance.
Giving helpful feedback	http://www.sussex.ac.uk/tldu/ideas/eval/ceq http://celt.ust.hk/instr/instr11.htm McGrath , Yeung, Comfort & McMillan (2005) Madden, Dillon & Leak (2010)	The feedback of our teacher helps me to improve myself.