# ASSESSING ACADEMIC WRITING SKILLS

# IN TURKISH AS A FOREIGN LANGUAGE

FATMA KÜÇÜK

BOĞAZİÇİ UNIVERSITY

2017

# ASSESSING ACADEMIC WRITING SKILLS
# IN TURKISH AS A FOREIGN LANGUAGE

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Fatma Küçük

Boğaziçi University

2017

Assessing Academic Writing Skills in Turkish as a Foreign Language

The thesis of Fatma Küçük

has been approved by:

Assist. Prof. Aylin Ünaldı<br>
(Thesis Advisor) ________________________

Assoc. Prof. Gülcan Erçetin ________________________

Assist. Prof. Sevdeğer Çeçen<br>
(External Member) ________________________

January 2017

# DECLARATION OF ORIGINALITY

I, Fatma Küçük, certify that

- I am the sole author of the thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature ...................................

Date ...17.02.2017.....

# ABSTRACT

## Assessing Academic Writing Skills in Turkish as a Foreign Language

The purpose of this study is to present validity evidence from multiple perspectives for the newly developed test of writing for learners of Turkish as a Foreign Language (TFL). Following Weir’s (2005) socio-cognitive framework, the study provides evidence for scoring validity along with cognitive and context validity aspects of the TFL writing test. The test aims to assess academic writing proficiency of foreign students, mostly Exchange and Erasmus, at Boğaziçi University. Two different task types – graph interpretation and essay tasks – were developed drawing on the level descriptors of the Common European Framework of Reference (Council of Europe, 2001). The test was administered to 47 students registered in intermediate and advanced Turkish for Foreigners classes (TKF) at Boğaziçi University. The test scores were analyzed through many-facet Rasch measurement model (MFRM) to investigate task difficulty, rating scale effectiveness and possible involvement of rater effects. Generalizability theory analysis was also carried out to investigate dependability of test scores. The two task types were also analyzed qualitatively to examine differing cognitive and linguistic demands placed by the tasks along with a posteriori analysis of the test scores to examine the effect of the task types on student performance. The findings from these analyses provided substantial support for the validity of the TFL writing test. The study also highlighted the possible differential effects of two task types on the performance of test takers and provided explanatory discussion as to the reasons of this.

# ÖZET

## Yabancı Dil Olarak Türkçede Akademik Yazma Becerilerinin Değerlendirilmesi

Bu çalışmanın amacı, Türkçeyi yabancı dil olarak öğrenenler için yeni geliştirilen yazma sınavının geçerliliğine farklı yönlerden kanıt sunmaktır. Çalışma, Weir'in (2005) sosyo-bilişsel çerçevesini temel alarak puanlama, bilişsel ve bağlamsal geçerliliği yönlerinden sınavın geçerliliğine kanıt sağlamıştır. Yazma testinin amacı Boğaziçi Üniversitesi'ne gelen yabancı öğrencilerin akademik yazma yeterliliklerini ölçmektir. Bu sebeple, Avrupa Ortak Dil Çerçevesi (CEFR, Council of Europe, 2001) seviye tanımlayıcılarından yararlanarak iki farklı yazma görevi geliştirildi – grafik yorumu ve düz yazı. Bu görevler, Boğaziçi Üniversitesi orta ve ileri seviye yabancılara Türkçe derslerine (TKF) kayıtlı 47 öğrenciye uygulandı. Elde edilen puanlar, sınavların zorluğunu, değerlendirme ölçeğinin yeterliliğini ve olası değerlendiren etkilerini incelemek amacıyla çok yönlü Rasch ölçüm modeli (MFRM) kullanılarak analiz edildi. Sınav puanlarının güvenilirliğini ölçmek için Genellenebilirlik kuramı da kullanıldı. Testte kullanılan iki farklı sınav görevinin öğrenci performansı üzerindeki etkisini incelemek için ise, elde edilen sınav puanlarının incelenmesinin yanı sıra, bu sınav görevleri bilişsel ve dilbilimsel özelliklerindeki farklılıkları açısından nitel olarak incelendi. Bu incelemelerden elde edilen bulgular, yabancılar için geliştirilen Türkçe yazma testinin geçerliliği konusunda önemli kanıtlar sunmuştur. Bu çalışma, ayrıca iki tür sınav görevinin öğrencilerin performanslarında yapabileceği farklı etkiler konusunun altını çizmiş ve bunun nedenleri üzerine açıklayıcı bir tartışma sunmuştur.

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, Assist. Prof. Aylin Ünaldı, for her excellent support, guidance and encouragement throughout this study. Her contribution to this thesis and me personally is too precious to be expressed with words. I also owe many thanks to the rest of my thesis committee, Assoc. Prof. Gülcan Erçetin and Assist. Prof. Sevdeğer Çeçen, for their encouraging comments and constructive feedback. Special thanks to Assoc. Prof. Gülcan Erçetin for her crucial role in building the background of this study.

I wish to send my heartfelt thanks to Prof. Ayşe Gürel for helping me with the data collection and providing illuminating feedback whenever I needed it. I am thankful to Assist. Prof. Bengü Börkan for teaching me how to conduct and interpret many-facet Rasch analysis. I am truly grateful to Assoc. Prof. Carol Myford, who responded to all my questions patiently and sincerely as if I was her own student during the challenging times I had with the many-facet Rasch analysis. Special thanks should go to my friends Talip Gülle and Yavuz Kurt for their contributions in data collection and rating. Their friendship and support throughout my MA years were invaluable. I feel indebted to Talip Gülle for being ready to help me at all times.

Above all, I wholeheartedly thank my husband, Serkan Üçpınar, for bearing with me during all the hardships I went through. I could have never accomplished this without his emotional support, endless trust and motivation throughout my years of MA study. Finally, I would like to thank my beloved parents, Sabit and Semra Küçük, who have always encouraged me to continue my education and supported me with their unconditional love and endless care. I feel very lucky to be a part of this family.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction to the study

Test usefulness is argued to be the most important consideration in language test construction (Bachman & Palmer, 1996). The model of test usefulness proposed by Bachman and Palmer (1996) involves six qualities, two of which are considered to be critical in evaluating test usefulness: validity and reliability. Reliability and validity are also accepted as essential measurement qualities as these qualities are often used to justify the use of test scores in making inferences about one's abilities or to make decisions (Bachman & Palmer, 1996). Reliability is about "consistency of measurement across different characteristics or facets of a testing situation, such as different prompts and different raters" (Weigle, 2002, p. 49). Reliability is an essential quality in that consistent test scores are needed in order to rely on the inferences and decisions that are made based on these scores. Validity is, on the other hand, concerned with "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman & Palmer, 1996:21). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p.5). It can be inferred from Messick's definition of validity that certain kind of evidence needs to be collected in order to claim that interpretations and actions based on test scores are appropriate and meaningful. In the field of second language assessment several validity frameworks that can guide test development and validation have been proposed such as Messick's unified framework of validity

(1989), Kane’s interpretative argument framework (1992, as cited in Bachman, 2004) and Weir’s socio-cognitive validity framework (2005). Weir (2005) suggests that “the more comprehensive the approach to validation ... the more secure we can be in our arguments for the validity of a test” (p.47). Therefore, Weir (2005) offers a comprehensive framework of validity, which encompasses a priori validation components of context and cognitive validity and a posteriori validation components of scoring, consequential and criterion-related validity. A full-fledged validation study would attempt to gather evidence from these multiple perspectives and each piece of evidence would help to improve the understanding of the usefulness of the test. In addition, validity and reliability are no longer seen as different test qualities in this framework, but as part of a unified approach to validity as in Messick’s (1989) approach. Weir’s (2005) framework has been used for the validation of Cambridge ESOL examinations by various scholars: Shaw and Weir (2007) for validation of writing tests, Khalifa and Weir (2009) for validation of reading tests, Taylor (2011) for validation of speaking tests. With respect to writing assessment, Shaw and Weir (2007) state that three major components of validity framework (i.e., cognitive, context and scoring validity) constitute the core of test validation as these components make it possible to provide essential evidence to support validity claims. For this reason, the present study, which focuses on the validation of a newly developed writing test, focuses mainly on scoring validity as well as aspects of cognitive and context validity.

## 1.2 Aims of the present study and the research questions

The aim of the current study is basically to investigate and provide theoretical and empirical evidence for validity and reliability of the two writing tasks and related rubrics developed to assess academic writing proficiency of learners of Turkish as a foreign language (TFL). The content of the tasks was informed by the CEFR level descriptors, proposed specifically for B2 and C1 levels. The test was designed with the aim of eliciting the type of abilities that test takers need in academic settings. The target population of the test is foreign students, mainly the Exchange or Erasmus students, coming from various countries to study at Boğaziçi University. These students are offered Turkish for Foreigners (TKF) courses by the department of Turkish Language and Literature at different proficiency levels. Each semester nearly 150 students enroll in these courses (Gülle, 2015). These students are placed in the classes based on their own perceptions of their Turkish proficiency or the judgments of the course instructors. Clearly, there is a need for a standardized test based on which meaningful and appropriate interpretations can be made regarding students' proficiency, and decisions about which classes they need to be placed. The current test of writing was developed as a part of a larger project, which includes reading, speaking and listening tests with the aim of responding to this need. The writing test involves two tasks of different genres each of which is assumed to tap on academic writing skill; information transfer task (graph interpretation) and an essay task. This study was designed to provide evidence especially to the scoring validity of this writing test as well as providing arguments for the cognitive and context validity of the test.

In line with the above stated aim of the study, the following questions were investigated:

1. How do the graph interpretation and essay tasks differ from each other in terms of cognitive and linguistic demands?

   1.1 What are the qualitative differences between the tasks in terms of cognitive and contextual features?

   1.2 Do the students perform differently on these task types?

   1.3 Are the tasks different in terms of difficulty?

   1.4 Do the tasks reliably separate students into distinct proficiency categories?

2. How reliably does the rating scale function within this particular group of raters?

3. To what extent is the quality of ratings influenced by rater effects?

4. How dependable are the scores assigned to the examinees?

   4.1 What is the relative contribution of persons, tasks and raters to the overall variability in the ratings?

   4.2 How many writing tasks and raters are necessary to achieve /increase acceptable levels of reliability?

The first research question is concerned with the cognitive and context validity. Two different types of tasks are investigated in terms of their cognitive demands and contextual features. The main concern of this question is to discuss how cognitive and contextual variables that might impinge on performance differ across tasks and how these actually result in differential performance in the tasks in the test under scrutiny. The second, third and fourth questions are related with aspects of scoring validity. The main concern of the second question has to do with the reliability of the rating scale that was used to rate students' responses. The third question seeks evidence for involvement of any considerable rater effects in the test scores since variability in scores associated with rater characteristics are considered to be one type of construct irrelevant variance that adversely affect scoring validity. The fourth question investigates the dependability of test scores by examining main effects of different sources of variability (i.e., raters and tasks) through generalizability theory analysis.

## 1.3 Overview of the thesis

Following this introduction chapter, the chapters of the thesis is organized as follows: Chapter 2 presents a review of literature on some of the basic considerations in the construction of language tests including validity and reliability issues, Weir’s (2005) framework for developing and validating writing tests, an overview of theoretical models of writing, issues on scale development, and two measurement models used to identify sources of variances involved in performance test scores (many-facet Rasch measurement model and Generalizability theory). Chapter 3 reports the methods and procedures followed in the study. Detailed information about the aims of the study, the research questions, participants, instruments, data collection procedures, scoring procedures and data analysis is provided in that chapter. Chapter 4 presents the results concerning four research questions. The chapter includes qualitative analysis of the test tasks in terms of cognitive and contextual parameters, and the results of quantitative analysis used to investigate task difficulty, rating scale reliability, rater effects and score dependability. A detailed discussion of the results in relation to the research questions are presented in Chapter 5. Finally, chapter 6 provides a summary of the findings, research implications, limitations of the current study and suggestions for future research.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

This chapter intends to present a review of literature on basic considerations in constructing a language test. Regarding this issue, Bachman and Palmer (1996, p. 17) argue that the most important quality of a test is its usefulness. In order to evaluate test usefulness, Bachman and Palmer (1996) suggested a model of test usefulness which includes six qualities of a test: reliability, construct validity, authenticity, interactiveness, impact, and practicality (p. 17). The chapter starts with a brief discussion of these qualities specifically focusing on reliability and validity aspects as these two qualities are considered critical for language tests in that using test scores to make inferences or decisions is justified through the establishment of these qualities (Bachman & Palmer, 1996). Then, the framework offered by Weir (2005) for developing and validating writing tests is introduced, and relevant issues regarding cognitive, context and scoring validity components are discussed while the consequential and criterion-related validity are not included as they are beyond the scope of this study. This framework which is also operationalized by Shaw and Weir (2007) in validating the writing component of Cambridge ESOL examination is considered important in that it provides validation schemes for both priori and posteriori evidence on the validation of the current test. The chapter includes definition of writing ability as a construct and an overview of theoretical models which shed light on identifying the construct. It then discusses issues on writing task characteristics and scale development which are considered to influence the reliability of scores, and thus validity of the test. The chapter closes with an introduction of two measurement models that are frequently used in performance assessment to identify sources of error variances involved in test scores: many-facet Rasch measurement and generalizability theory.

## 2.2 Test usefulness

Bachman and Palmer (1996, p. 17) consider test usefulness as the most important consideration in designing and developing a language test. With the model of test usefulness that they propose, Bachman and Palmer attempt to answer the question, “How useful is this particular test for its intended purpose?”(p. 17). The model of test usefulness consists of six qualities; reliability, construct validity, authenticity, interactiveness, impact and practicality. Bachman and Palmer suggest that rather than trying to maximize all these qualities in a test, which is hardly possible, it is more appropriate for test developers to find a balance among these qualities depending on the test purpose, particular test takers and target language use domain. To illustrate, a test designer who aims to design a large-scale test to be used for important decisions is likely to desire highest levels of reliability and validity whereas a teacher may want to prioritize authenticity, interactiveness and impact qualities in a classroom test (Bachman & Palmer, 1996). Bachman and Palmer (1996) consider reliability and validity as being critical for tests since these qualities are utilized to justify the use of test scores as the indicators of the ability intended to be measured. Therefore, these two qualities will be the focus of the present study and the other qualities will be mentioned when necessary in relation to validity and reliability as these qualities complement each other in the overall usefulness of a test.

Weigle (2002) defines reliability in relation to the assessment of writing as “consistency of measurement across different characteristics or facets of a testing situation, such as different prompts and different raters" (p. 49). In other words, a test is claimed to be reliable if test takers are rewarded with the same score from different versions of a test, by different raters or on different occasions (Weigle, 2002). Reliability is thought to be an essential quality of test scores because if consistent results are not obtained, one cannot rely on the inferences and decisions based on them (Bachman & Palmer, 1996). Bachman and Palmer maintain that although it is virtually impossible to eliminate all potential inconsistencies, it is possible to minimize those inconsistencies through test design by minimizing variation in task characteristics. With regard to writing tests, reliability is said to be affected both by variables related to writing task characteristics such as the topic, discourse mode, the number of tasks test taker is asked to respond; and by variables related to the scoring process including the type of rating scale, rater characteristics, and rater training (Weigle, 2002). These issues will be detailed later in this chapter.

Construct validity is defined as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman & Palmer, 1996: 21). Messick (1989, as cited in Messick 1996) defines construct validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 6). Messick's influential definition implies that to be able to justify that a specific score interpretation is meaningful and appropriate, evidence is needed. To collect such evidence, one first needs to define the construct to be measured. Construct refers to "the specific definition of an ability that provides the basis for a given test or test task for interpreting scores derived from this task" (Bachman & Palmer, 1996, p. 21). Construct validity is therefore about the extent to which a particular interpretation is justified in relation to the construct (ability) measured. Bachman and Palmer state that construct validity is also related with the generalization of score interpretations to the domain of generalization, which is described as “the set of tasks in the target language use domain (TLU) to which the test tasks correspond” (p. 21). The latter aspect of construct validity suggests that the interpretations about language ability should be extended beyond testing situation to a specific TLU domain. Moreover, the second aspect of construct validity is closely related with authenticity which is described as “the degree of correspondence of the characteristics of a given task to the features of a TLU task (Bachman & Palmer, 1996, p. 23).

Authenticity is considered to be an important quality as it is directly linked to construct validity in that the generalization of score interpretations beyond the testing situation to the TLU domain is determined by the degree of authenticity of the test tasks (Bachman & Palmer, 1996). Weigle (2002) argues that achieving authenticity may not be easy in certain cases such as in English for Academic Purposes (EAP), one of the widest areas in terms of testing writing. An example Weigle gives is a typical writing test which includes a timed impromptu essay on a topic on which test takers have no prior knowledge. She maintains that such a task suffers from inauthenticity to some extent and lists four reasons, one of which is that academic writing is typically actualized by using source materials such as assigned readings, lecture notes, or class discussions while in testing situations using such source materials is not possible. Messick (1994) claims that such kind of factors that jeopardize authenticity lead to construct under-representation, which he considers as a threat to construct validity. Construct irrelevance is another threat to construct validity which jeopardizes the directness of a test. With the term direct assessment, Messick (1994) implies open-ended tasks and judgmental scoring that are utilized with the aim of preventing test takers from involving in any construct irrelevant factors such as “testwiseness in coping with various item types, differential tendencies towards guessing, and other artificial restrictions on students’ representations of problems or on their modes of thinking or response” (p. 21). Messick (1994) maintains that authenticity and directness are two concepts related with performance tests which are typically employed in the assessment of productive skills (speaking and writing); therefore construct under-representation and construct irrelevance lead to concerns specifically in performance tests. In order to minimize these threats, the most practical suggestion would be not to leave any relevant skill out, and not to include any irrelevant factors in the course of test development (Messick, 1994).

Construct validation is seen as an on-going process of collecting evidence to justify a particular interpretation of test scores (Bachman & Palmer, 1996). At this point, an important question to be asked is what kinds of evidence can be collected to demonstrate that the result of a particular test is actually a true indicator of the underlying ability (construct). A useful framework that can guide a validation study in its necessary steps is the socio-cognitive framework proposed by Weir (2005). In this study, the guidance will be taken from the socio-cognitive framework in attempts to justify the validity of the writing test developed to assess academic writing skills of learners of Turkish as a Foreign Language (TFL).

## 2.3 The socio-cognitive validity framework

Weir (2005) presents a comprehensive validity framework to account for different types of evidence that can be collected at different stages during test development and post implementation phase within an evidence-based paradigm. Before examining this framework, several things should be clarified in order to comprehend and evaluate it better. First, although construct validity is often used as a superordinate term that includes all other aspects of validity, Weir (2005) prefers validity as the superordinate term, both of which refer to the same concept.  Second, although reliability and validity are traditionally viewed as different test qualities (polarized), Weir (2005) regards reliability as one type of validity evidence, and uses the term scoring validity instead.

The validity framework offered by Weir (2005) is comprehensive because it involves both a priori evidence which should be collected before the test administration and a posteriori evidence that should be collected after the test is administered. It is claimed that "the more comprehensive the approach to validation, the more secure we can be in our claims for the validity of a test" (Weir, 2005, p. 47). Weir provides an overview of this framework for four macro-skills of reading, listening, speaking and writing. Since this paper is basically concerned with writing skill, the framework that is offered specifically for writing skill and operationalized by Shaw and Weir (2007) to validate Cambridge ESOL examinations will be investigated.  Figure 1 illustrates the framework.

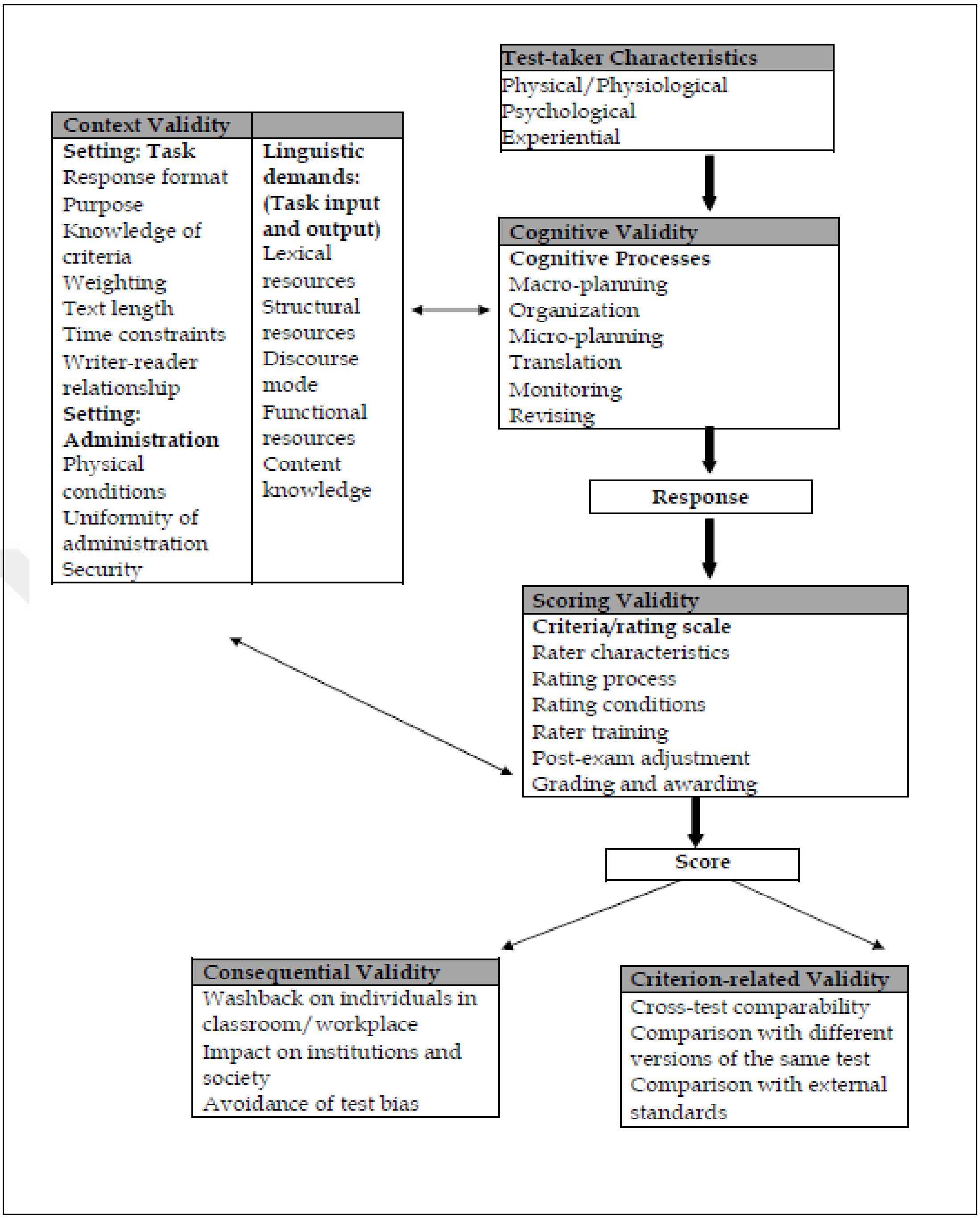


Figure 1. A socio-cognitive framework for validating writing tests (Weir, 2005, p. 47)

The framework is called socio-cognitive because the act of writing is viewed as a cognitive activity that takes place within a context of use (Weir, 2005). The arrows in Figure 1 show “the principal direction(s) of any hypothesized relationships” between the components (p. 43). The components are temporally conceptualized from the top to the bottom, which means that it depicts a map of the kind of evidence that should be collected at each stage before and after test event (Weir, 2005). Context and theory-based validity form a priori validation components; and scoring validity, consequential validity and criterion- related validity constitute a posteriori validation components. Shaw and Weir (2007, p. 10) argue that the three components of the framework are critical in any language testing activity: 1) test takers' cognitive abilities, 2) the context in which the task is performed, 3) the scoring process. These three components refer to cognitive validity, context validity and scoring validity respectively. It is claimed that a comprehensive analysis of testing process with respect to these three dimensions makes it possible to "provide theoretical, logical and empirical evidence to support validity claims and arguments about the quality and usefulness of writing tests" (Shaw and Weir, 2007, p. 11). Therefore, these three dimensions (called internal dimensions by Shaw and Weir) will constitute the basis of the present study while the consequential and criterion-related validity will not be included.

### 2.3.1 Theory-based validity

Theory-based validity defined in this framework is mainly related to cognitive aspect of the act of writing, which is why this component is refined as cognitive validity by Shaw and Weir (2007). Shaw and Weir (2007) describe the cognitive validity of a writing task as "a measure of how closely it represents the cognitive processing involved in writing contexts beyond the test itself" (p. 34). This definition suggests that the cognitive processes activated by a particular test should reflect the processes that we are involved in engaging the real life writing tasks. An important question to be enquired at this point is what kind of cognitive processes are employed during the act of writing. Several theoretical models of writing ability have been proposed with the aim of explaining various cognitive processes that take place during the act of writing (i.e. Flower & Hayes, 1981; Breiter & Scardamalia, 1987; Grabe & Kaplan, 1996; Hayes, 1996; Kellogg, 1994; Field, 2004; as cited in Shaw and Weir, 2007). Although most of these models draw basically on the processes of the first-language writing (Weigle, 2002), they are of great value in understanding the nature of writing ability in a foreign language.

An influential model was proposed by Flower and Hayes (1981) to account for writing process. An illustration of Flower and Hayes Model is presented in Figure 2.

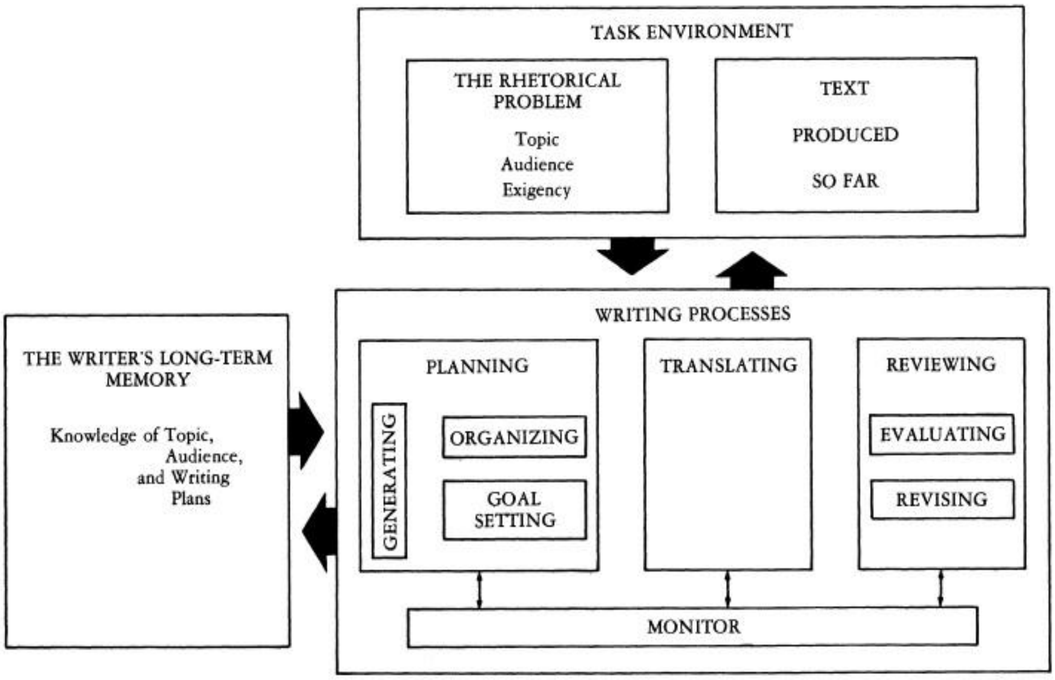


Figure 2. The Flower and Hayes writing process model (1981)

As can be seen from Figure 2, the model divides the writing process into three major components and the largest space belongs to writing processes, which implies that they are naturally at the core in the act of writing. Writing processes consist of planning, translating, reviewing, which are all controlled by monitor simultaneously (Flower & Hayes, 1981). Ideas are generated and organized in planning stage, and the goals set by the writer guide the creation of ideas. The ideas generated in the planning stage in an abstract form are transcribed into written language in the translation stage. Possible mismatches or gaps between ideas and written language are evaluated; revisions are made to minimize the gaps during the reviewing stage. As the text is being composed, the current process and progress are actively monitored by the writer (Flower & Hayes, 1981). These cognitive processes are influenced by the task environment which consists of writing assignment and text produced so far; and by the writer's long term memory which includes writer's knowledge of topic, audience and stored writing plans (Hyland, 2002). Although this early model has been influential, it also received criticism because of the fact that it fails to explain different processing strategies that are likely to be employed by writers with varying language abilities (Grabe and Kaplan, 1996). In other words, it does not differentiate between skilled and less skilled writers.  Breiter and Scardamalia (1987, as cited in Grabe and Kaplan, 1996) proposed two models of knowledge telling and knowledge transforming in order to capture different types of composing behavior among skilled and less-skilled writers. The knowledge telling model basically suggests that:

> Novice (less-skilled) writers plan less often than experts, revise less often and less extensively, and are primarily concerned with generating content from their internal resources. Their main goal is simply to tell what they can remember based on the assignment, the topic, or the genre. (Breteir and Scardamalia, 1987; as cited in Hyland, 2002, p. 28)

These types of strategies are claimed to be useful when writing about personal experiences or feelings, and telling narratives, the processing demands of which are relatively simple and do not require much planning (Grabe & Kaplan, 1996). However, when it comes to writing tasks which require more demanding processing, the knowledge-telling model provides little help, and the knowledge transforming model comes into play. The knowledge transforming model describes:

> How skilled writers use the writing task to analyze problems and set goals. These writers are able to reflect on the complexities of the task and resolve problems of content space and a rhetorical space, so that there is continuous interaction between developing knowledge and text. Knowledge transforming thus involves actively reworking thoughts so that in the process not only text, but also ideas may be changed. (Breiter and Scardamalia, 1987; as cited in Hyland 2002, p. 28)

The explanation cited above indicates that the knowledge transforming model differs from the knowledge telling model in that it involves problem analysis, goal setting, problem solving processes and creation of new knowledge, and the knowledge telling is actually one component of the knowledge transforming model (Grabe & Kaplan, 1996). The two- model process proposed by Breiter and Scardamalia (1997, as cited in Grabe & Kaplan, 1996) is important because it not only differentiates between expert and novice writers, but it also explains why writing tasks may differ in difficulty, even for equally skilled writers. Accordingly, task difficulty will differ from person to person based on background knowledge on the topic and genre familiarity (Grabe & Kaplan, 1996).

Weir (2005) argues that although these two major models of writing process have led to a significant improvement on the view of writing, they fail to account for the contextual factors that affect writing process. The model of writing as communicative language use proposed by Grabe and Kaplan (1996) was one of the first model incorporating internal processing and contextual factors (Weir, 2005). This model was adapted from Chapelle et al.'s (1993, as cited in Grabe and Kaplan, 1996) model of communicative language use developed for academic language performance in listening, speaking, reading and writing. As shown in Figure 3, the model basically consists of a context and language user's verbal working memory which have several categories. Two major components of context (situation and performance) with their subcomponents constitute the external social context of writing situation.

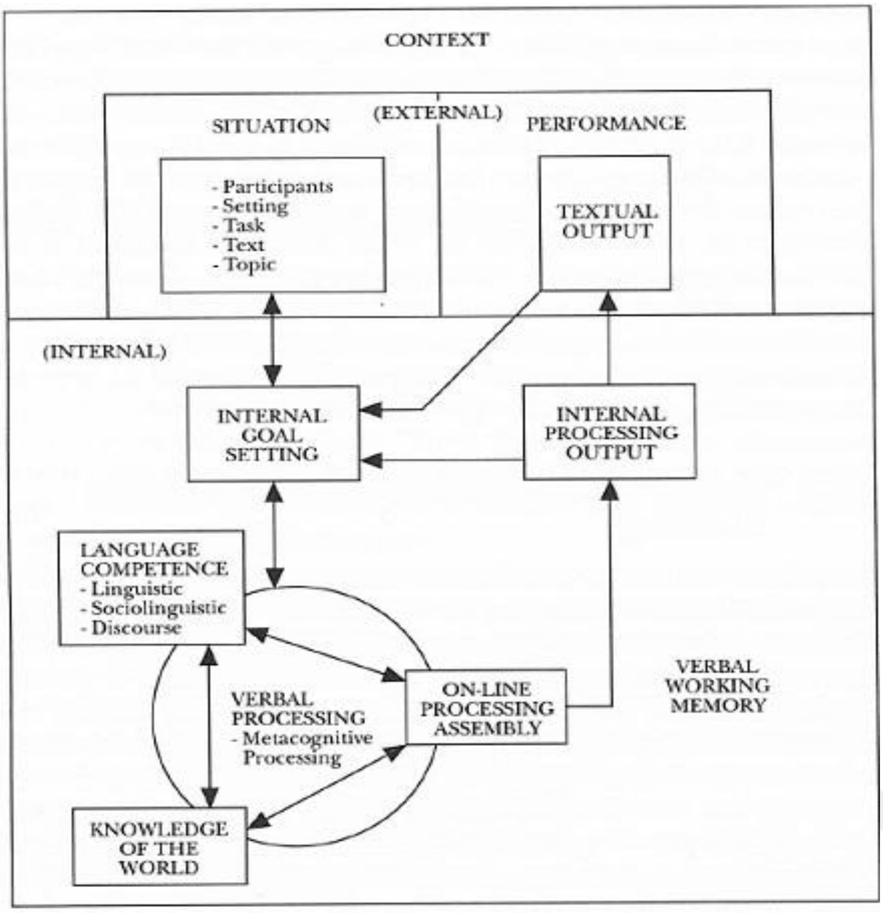


Figure 3. Model of writing as communicative language use (Grabe & Kaplan, 1996)

Issues regarding anticipated register constrains, genre constrains, and communicative functions, norms and conventions are part of the situation component. The other main component of the model is the verbal working memory which includes processing activities of the language user during the act of writing. It consists of three subcategories: internal goal setting, verbal processing, and internal processing output. Verbal processing itself has three subcategories which are language competence, knowledge of the world, and on-line processing assembly. Language competence is composed of three categories (linguistic knowledge, discourse knowledge, sociolinguistic knowledge) drawing upon the work of Hymes (1972, as cited in Grabe & Kaplan, 1996), Canale and Swain (1980, as cited in

Grabe & Kaplan, 1996) and Bachman (1990, as cited in Grabe & Kaplan, 1996). All these categories are considered to activate each other in a cyclical way during the execution of a writing task.

It can be concluded from the models listed above that Flower and Hayes model (1981) basically focused on the cognitive aspects of writing whereas Breiter and Scardamalia (1987, as cited in Grabe & Kaplan, 1996) differentiated between the cognitive processes that expert and novice writers employ. Grabe and Kaplan (1996), on the other hand, emphasized contextual factors besides cognitive processes and included language competence, which is an essential component especially in second language writing. Although these three models proposed by Flower and Hayes (1981), Breiter and Scardamalia (1987, as cited in Grabe & Kaplan, 1996), Grabe and Kaplan (1996) suggested different components to account for the nature of writing and seem to view writing in somewhat distinct ways, they all indicate that writing is a complex process that involve several cognitive processes, and influenced by factors beyond the writer himself such as topic, audience, setting, participants etc.

These theoretical models may be useful in defining “writing” as a construct as well as determining task characteristics. To illustrate, Shaw and Weir (2007) state that knowledge telling and knowledge transforming model proposed by Breiter and Scardamalia (1987, Grabe & Kaplan, 1996)) is helpful in determining lower and higher level tasks for Cambridge ESOL examinations (KET, PET, FCE, CAE and CPE). Accordingly, narrative or instructional texts aimed for lower levels are considered to require knowledge telling skills whereas argumentative texts at higher levels entail knowledge transforming skills.

Before attempting to define writing as a construct to be assessed, it is equally important to emphasize the social aspects of writing and introduce relevant theoretical models as the act of writing is often carried out within a social context to achieve a particular purpose intended for particular audience (Hamp-Lyons and Kroll, 1997; as cited in Weigle 2002). Hayes (1996:5, as cited in Weigle, 2002) argues that "what we write, how we write, and who we write to is shaped by social convention and by our history of social interaction" (p. 19). Theories of communicative language ability might be helpful in conceptualizing writing ability as a socially embedded task (i.e., Canale and Swain, 1980; Bachman, 1990). A comprehensive model of communicative language ability which was proposed by Bachman (1990) and improved by Bachman and Palmer (1996) consists of two broad areas: language knowledge and strategic competence. Organizational knowledge and pragmatic knowledge constitute two main components of language knowledge. Organizational knowledge includes grammatical knowledge (knowledge of vocabulary, morphology, syntax and phonology), textual knowledge (knowledge of how the units of grammatical knowledge are brought together to create coherent texts). Pragmatic knowledge involves functional knowledge (knowledge of how different communicative functions are accomplished through language) and sociolinguistic knowledge (knowledge of how to use language in various social context in an appropriate way). Strategic competence which enables individuals to involve in goal setting, assessment and planning is described as "a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities" (Bachman & Palmer, 1996, p. 70). It is also pointed out that topical knowledge, affective schemata and personal characteristics of individuals interact with language knowledge and strategic competence in language use (Bachman & Palmer, 1996).

The Common European Framework (CEFR), which has become one of the key reference documents in second language assessment as well as learning and teaching practices since its publication (Harsch & Rupp, 2011), delineates a model of language functions based on Bachman and Palmer's model of communicative language ability. This model consists of linguistic competence, sociolinguistic competence and pragmatic competence. Linguistic competence comprises lexical, phonological, grammatical, semantic, orthographic and orthoepic knowledge. Sociolinguistic knowledge is associated with knowledge and skills for proper language use in social contexts including elements of language that address social relationships, conventions of appropriate behavior, and varieties in register, dialects and accent. Finally, pragmatic competence encompasses discourse and pragmatic competence. Discourse competence is concerned with the ability of ordering sentences or utterances to form coherent pieces of language. Functional competence is about using appropriate language to engage in communicative functions, and design competence. However, strategic competence is not a component of communicative language competence model in the CEFR though communicative language strategies such as planning, execution, monitoring and repair action are claimed to be inevitable while carrying out communicative tasks in the CEFR.

These models of communicative language ability which were mentioned above indicate that using a language (written or spoken form) entails different types of knowledge besides linguistic competence. In fact, being able to use linguistic or language knowledge properly in various social contexts in accordance with the norms and conventions of a particular society to achieve communicative functions is strongly emphasized in these models. Therefore, writing is delineated as a socially embedded skill.

#### 2.3.1.1 Defining writing as a construct to be assessed

The difficulty of developing a clear statement of writing as the target construct has been emphasized in the literature (i.e., Cumming, Kantor, Powers, Santos, & Taylor, 2000) just as in Purve's statement: "Logically and empirically, there seems little merit in references to a unitary construct of writing" (1992, as cited in Cumming et al., 2000, p. 2). Weigle (2002) also points out that the ability that one intends to measure might differ for each testing situation depending on test purpose, test takers and target language use situation. However, as stated at the beginning of the chapter, developing a clear definition of the ability to be measured is crucial to be able to justify the interpretations and decisions made based on the test score.  In the lights of the theoretical models that account for the nature of writing and communicative language ability as well as consideration of the academic settings as target language use domain and university students being target test takers, writing ability can be described as "generating and organizing ideas in accordance with the purpose, audience, and genre of the assigned task to produce coherent and appropriate pieces of text." It should be emphasized here that this definition is aimed for academic writing rather than its general conception. The concept of academic writing is used for writing that takes place within university context which has its own expectations for genre and conventions established by academic discourse community (Spack, 1988).

Having discussed the theoretical (cognitive and communicative) models in an effort to comprehend the nature of writing ability and  develop a clear construct definition, the next section will be dedicated to the second component of Weir's socio-cognitive framework: context validity.

### 2.3.2 Context validity

Context validity, which is often termed as content validity (i.e., Fulcher & Davidson, 2007), refers to “the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample” (Weir, 2005, p. 19). This definition implies that task characteristics and settings of tasks should reflect “performance conditions of the real-life context” as much as possible (Shaw & Weir, 2007, p. 63). Context validity embodies linguistic and content requirements of the tasks that are necessary for a satisfactory task completion, and the features of the task and administration setting as shown in Figure 4.

Setting: Task

- Response Format
- Purpose
- Known criteria
- Weighting
- Order of items
- Time Constrains

Setting: Administration

- Physical conditions
- Uniformity of administration
- Security

Linguistic demands:
Task input and output

- Discourse mode
- Channel
- Text length
- Writer-reader relationship
- Nature of Information
- Functional resources
- Structural resources
- Lexical resources
- Content knowledge

Figure 4. Aspects of context validity (Weir, 2005)

Weir (2005) points out that there is a symbiotic relationship between context and cognitive validity, which means that they are both influenced by each other. Any decisions about task characteristics may affect the cognitive processes that are activated by the task, and therefore the quality of the output. In the following section, instead of discussing every aspect of context validity identified by Weir, several task dimensions that are listed by Weigle (2002) will be explored with the discussion from related literature for the purpose of the present study. While the task dimensions provided by Weigle (2002) are closely related with the aspects of context validity listed by Weir (2005), not all of the aspects listed in Figure 4 will be covered. In this section, task features that influence task difficulty, the use of the CEFR as the reference point to determine the task content, and task specification issues will be discussed as they are mostly relevant to the context validity arguments in the present study.

Weigle (2002) underlines the salient components of task dimensions with reference to Hout's (1990) review of literature on direct writing assessment. These are discussed in the following sections.

#### 2.3.2.1 Discourse mode

Weigle (2002) argues that investigating the role of discourse mode in assessment of writing is demanding as it includes various dimensions which may have an effect on writing quality individually or together. Discourse mode basically involves categories of genre, rhetorical task, patterns of exposition and cognitive demands. Weigle (2002) defines genre as "the expected form and communicative function of the written product; for example, a letter, an essay, or a laboratory report" (p. 62). Rhetorical tasks refer to such types of discourse modes as narration, description, exposition and argument. Pattern of exposition can be considered as subcategories of exposition such as comparing and contrasting, discussing causes and effects or advantages and disadvantages of a given subject. Previous research examining the effects of discourse mode on writing performance demonstrated contradictory results. Crowhurst (1980, as cited in Hout, 1990) observed syntactic variation between narrative and persuasive responses; students produced longer T-units in argumentative essays than in narratives. In line with these findings, Quellmalz, Capell, and Chou (1982, as cited in Hout, 1990) conducted a study with two hundred

11th and 12th grade high proficiency students and reported that expository tasks yielded better performance than narrative tasks. However, Reed (1992) found contradictory results in his study with 48 students who were of low, medium and high writing ability investigating the effect of modes of discourse (persuasive, descriptive, persuasive, narrative). It was found that students performed best on narrative tasks and worst on persuasive tasks while they performed moderately on descriptive and expository tasks regardless of their ability. In a more recent study, Lee and Kim (2007) found out that low intermediate level Korean college students differed in their performance across three rhetorical types of writing task. Students received the highest score on descriptive writing task, which was followed by narrative task, and the lowest on argumentative writing task. Their findings were in line with Way, Joiner and Seaman (2000)'s study in which they investigated the effects of descriptive, narrative and expository tasks on the writing quality of French learners by employing various evaluation methods (holistic scoring, length of product, mean lengths of T units and percentage of correct T units). It was found that descriptive task was the easiest of the three tasks and the expository task was the most difficult one.

Although studies on the effects of discourse mode yielded contradictory results, it can be concluded that tasks with different discourse mode tend to influence writing performance in various ways. Therefore, making inferences about students' ability by testing them on different types of tasks rather than one specific task may give a more comprehensive picture of their ability, and thus increases validity of the assessment process.

#### 2.3.2.2 Rhetorical specification and prompt wording

Rhetorical specification is about the amount of information provided in prompt with respect to audience, purpose, and the topic of a writing task (Brossell, 1986, as cited in Weigle 2002). Brossell (1983, as cited in Hout, 1990) investigated the effects of rhetorical specification on students' writing quality through three versions of six different topics. Those three versions varied according to extent of specification presented in the prompt. The first version provided only the topic itself. The second version included the topic, an introductory statement and a short instruction for a personal response. The third version was presented in a full specification with an elaborate scenario including the purpose and the audience. Contrary to the initial hypothesis which predicted that the prompt with full rhetorical specification would give rise to higher quality writing, Brossell found that moderately specified writing prompt elicited the highest quality responses, and he concluded that fully specified writing prompts were difficult to process and time consuming for examinees in a testing situation. Specification of audience has also been the focus of researchers although the results are mostly considered to be inconclusive. Most research on the effect of audience specification was conducted to investigate the degree of relationship between the writers and the audience rather than its existence or absence. In their study, Crowhurst and Piche (1979, as cited in Hout, 1990) found that students produced more linguistically complex sentences when they addressed audience with lower intimacy (i.e. their teachers) than when they addressed audience with higher intimacy (i.e. their friends).  Similarly, Craig's study (1988, as cited in Hout, 1990) revealed that students in sixth and eleventh grades used more objective and impersonal language in their essays when they wrote for a reader with high status (teachers) than when they wrote for a reader with low status (their friends).

Therefore, it can be assumed that the specification of the audience might be a factor that affects the quality of writing.

Studies related with prompt wording and structure seems to be inconclusive as far as it is known by the researcher. Brossell and Ash (1984, as cited in Hout) conducted a study to 900 college sophomores (native speakers of English) examining the effects of wording of writing prompts. They made use of writing assignments that differed in terms of the way they were presented. The assignments were presented either in a personal manner or neutral, and either as questions or commands. They found no statistical significance for any of the variables. Similarly, Osborne and Mulling (1994) examined students' preferences over the form of prompts and other factors regarding perceived difficulty of a topic. The prompts were presented in forms of questions or statements with 5- point likert scale. No statistically significant difference was found in students' choices as to different forms (questions or statements).

With regard to the wording of task instructions, Weir (2005) suggests that the wording of task rubrics which include all the instructions of what is required by the task needs to be unambiguous and accessible to all test takers so that no candidate will misinterpret the task. Moreover, the way that a prompt is worded will determine how candidates will respond to the task and what kind of cognitive strategies they will choose to complete the task. For example, a clear statement of purpose is likely to facilitate goal setting and monitoring strategies, which are two main cognitive strategies that are discussed in theoretical validity section (Weir, 2005). Therefore, a careful selection of prompt words and trialing the tasks to a small group of test takers before the administration might be helpful to make sure that task rubric is as clear, brief and simple as desired.

To summarize, findings related with rhetorical specification indicate that a medium level of task specification which is neither underspecified nor too detailed, along with clear instructions will help test takers to focus on the necessary information for a successful task completion.

### 2.3.2.3 Stimulus material

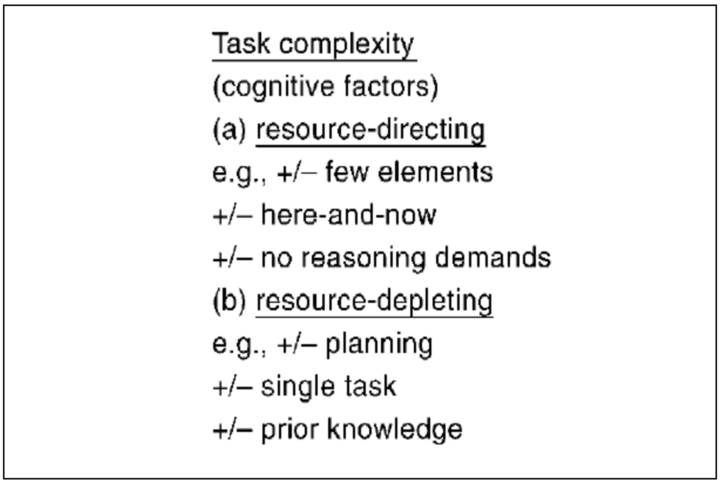
Stimulus material in a writing prompt involves any source material, such as a reading passage, a listening text, a quotation, a chart, or a picture (Weigle, 2002). Stimulus material is an important variable as writing tasks that are integrated with reading or listening tasks have gained popularity and used instead of or along with independent writing tasks to assess writing for academic purposes as evidenced in TOEFL IBT (Gebril & Plakans, 2009). The main argument for using source material in writing tasks is that it is considered to be an essential feature of academic writing in real life. In other words, integrated tasks are believed to increase the authenticity of academic writing tasks as students at colleges or universities are often required to write by integrating their knowledge from prior reading, lectures and other course materials (Cumming, 2013; Weigle, 2002). Besides, it is argued that giving source material provides all test takers the same piece of information to work with and activate their background schemata (Weigle, 2002). However, research has shown that providing stimulus material has several drawbacks as well as benefits. Lewkowicz (1997, as cited in Weigle, 2002) examined the effect of a background reading on writing quality of essays written by English learners in Hong Kong. Lewkowics found that although the reading text given as source material provided test takers with ideas, it did not influence the writing quality. What is more, those test takers who were provided with a background reading tended to borrow the language of the source text and could not develop their ideas as much as students who were not provided with a text. Unlike Lewkowics (1997), Plakans (2008) found several differences between writing-only-tasks and reading-to write tasks in test takers' planning stage, writing process and perception of task difficulty in his study with 10 non-native speakers of English. In the study, most of the participants preferred the reading–to-write task as the source texts provided them with context and ideas for writing. Writing-only tasks required more initial planning although reading-to-write tasks required more planning during writing process and displayed a more recursive process. Although most research focused on reading-writing tasks, Yang (2012) investigated examinees' academic writing strategies which they made use of on a graph –writing test administered to the learners of English in health science and medical majors. The results indicated that the task led the test takers to use strategies of graph comprehension, graph interpretation and graph translation which had a positive effect on their performance. Two thirds of the examines stated their preference for graph-writing tasks over reading-writing or writing only tasks for the fact that this type of tasks are more interesting, thought provoking, more similar to their academic assignments and provide clear-cut information. Qualitative analyses, however, showed a lack of graph-related lexical knowledge which the test takers considered as the main reason for task difficulty and several construct irrelevant sources of variance such as graph familiarity, topical knowledge and test wiseness strategies .

The studies investigating the effect of stimulus material on writing performance and cognitive processes show that tasks that are integrated with stimulus material may be helpful in providing test takers with content knowledge, which results in less initial planning despite the fact that it may be difficult to prevent textual borrowing. As opposed to more common practice of reading-to-write tasks or listening-to-write tasks, graph-writing tasks may be a feasible option to assess test takers' academic strategies such as graph interpretation and translation, which are likely to be employed in actual academic settings. However, involvement of possible construct irrelevant variances such as graph unfamiliarity, lack of graph-related lexical knowledge should be taken into consideration as they may lower validity of test scores.

#### 2.3.2.4 Task difficulty

Determining the difficulty level of tasks has to do with task dimensions, some of which are mentioned above as it is often achieved by manipulation of characteristics of task input (i.e., cognitive demands, discourse mode, length) and setting. In the literature (Skehan, 1998; Briendly, 1987 as cited in Robinson, 2001), it is suggested that difficulty level of the tasks increases with the extent of complexity of tasks, which involve explanation of abstract concepts or development of an argument compared with simpler tasks in which all the information and steps to complete the task is provided with concrete definitions. Besides, as the number of the elements to address and the amount of cognitive demands and reasoning to solve the task increase, task difficulty increases, too. On the other hand, factors such as familiarity and prior knowledge of tasks, topics and required cognitive operations, the amount of information provided and the precision of tasks decrease task difficulty. In the CEFR, it is stated the ease or difficulty of tasks cannot be predicted with certainty because each individual may approach to the same task differently depending on the strategies that they use, the degree of their language abilities available to cope with the task, individual characteristics, and specific conditions of the task administrations.

There are two influential models that attempt to explain task difficulty: Skehan and Foster (2001, as cited in Harsch & Rupp, 2011) proposed Limited Attention Capacity Model (LACM) to associate task complexity with the amount of attention a task requires from the learners. According to this model, the amount of attentional resources is limited, and increasing the level of task complexity will lead learners to shift their focus from linguistic form on meaning. This will result in linguistically less complex and more erroneous output. Robinson (2001) takes a different stance in his Multiple Attentional Resources Model, which is also known as Cognition Hypothesis claiming that learners posses different attentional resources and increasing task complexity will not lower the quality of their linguistic output, but may instead lead to higher level of linguistic complexity, lexical variation and accuracy. In his model, Robinson (2001) characterizes task complexity with several dimensions that can be divided into two main categories: resource-directing and resource- depleting. These dimensions are specified with + / − sign which shows their presence or absence as presented in Figure 5.



Task complexity
(cognitive factors)
(a) resource-directing
e.g., +/– few elements
+/– here-and-now
+/– no reasoning demands
(b) resource-depleting
e.g., +/– planning
+/– single task
+/– prior knowledge

Figure 5. Dimensions of task complexity (Robinson, 2001)

Robinson (2001) claims that resource-directing and resource depleting dimensions lead to a significant difference in the allocation of resource during L2 task performance. Accordingly, increasing task complexity along resource-directing dimensions, for example, by requiring test takers to use reasoning [+reasoning] or to consider many elements [−few elements] may direct learners' attention to certain linguistic features resulting in greater linguistic accuracy and complexity. On the other hand, making tasks more complex along with resource depleting dimensions, for example, by adding a secondary task to a primary task [−single task] or removing the planning time [−planning time] for task performance leads to greater demands on attention and working memory, and does not necessarily direct learners to any linguistic elements.

These two models have been empirically investigated by several researchers, and the results turned out to be inconsistent. To illustrate, Kuiken and Vedder (2008) attempted to put these two models to the test. In other words, they investigated the relationship between task complexity and linguistic performance of L2 learners in terms of syntactic complexity, lexical variation and accuracy. Their study with learners of Italian and French demonstrated that manipulation of task complexity by changing the number of the elements students had to address in letter writing tasks led to a significant increase in accuracy in more complex tasks but not syntactic complexity or lexical variety. However, Kuiken and Vedder (2007) reported that increasing task complexity improved lexical variation of learners' letter writing. Ishikawa (2006) investigated the effect of a different task complexity dimension on narrative writing quality by manipulating +/- Here and Now dimension. It was found that the narrative task that required students to write in past tense (-Here and Now) led to greater accuracy, fluency and syntactic complexity than the task that required students to write in present tense (+Here and Now).

Although these studies reported contradictory results, the general tendency of the findings suggest that increasing task complexity may have differential effects on writing quality. Therefore, it might be a good idea to prepare tasks that differ in difficulty to have a better understanding of examinees' writing ability.

#### 2.3.2.5 Aligning language examinations to the CEFR

Common European Framework of Reference has been increasingly used as the reference point in order to determine test content, task difficulty, and develop level specific or comparable tests. The CEFR was developed with the aim of providing "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001:1). However, today its influence has expanded beyond Europe, and many test developers feel the need to align the exams to the CEFR even though its authors state that the CEFR is not a how-to guide for constructing language tests (Khalifa, 2009). The CEFR offers a set of common reference levels that describe language proficiency ranging from the basic user (A1= Breakthrough, A2=Waystage) via the independent user (B1=Threshold, B2=Vantage), to the Proficient User (C1=Effective Operational Proficiency, C2= Mastery*)* along with empirically grounded illustrative descriptors of communicative activities, strategies and communicative language competences. Regarding its use in language assessment, the Framework suggests three basic ways:

1. For the specification of the content of tests and examinations.
2. For stating the criteria for the attainment of a learning objective, both in relation to the assessment of a particular spoken or written performance, and in relation to continuous teacher-peer-or self-assessment.
3. For describing the levels of proficiency in existing tests and examinations thus enabling comparisons to be made across different systems of qualifications.

(Council of Europe, 2001, p.19)

However, the CEFR is a descriptive framework rather than a practical assessment tool, and its use in test development has been criticized by some researchers. To illustrate, Weir (2005) argued that "in its present form the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing" (p. 281) as the CEFR is considered to fail in addressing aspects of validity at different levels of language ability; including the scales with incomplete contextual variables or performance conditions, little account of theory-based cognitive processes, and inconsistent and somewhat ambiguous wording of some descriptors. In order to overcome the limitations of the CEFR for developing writing tasks, members of the Association of Language Testers in Europe (ALTE) developed the CEFR Grid for the Analysis of Writing Tasks (Council of Europe, 2008) with the aim of facilitating the specification stage through analyzing test task content and other attributes. Besides, in order to facilitate the procedures of linking language tests to the CEFR, Council of Europe (2009) published the *Manual for Relating Language Examinations to the CEFR.* The Manual encompasses a comprehensive overview of basic considerations and suggests several steps to link examinations to the CEFR, and the suggested procedure has been applied in alignment of Cambridge ESOL examinations (Khalifa, 2009, Papp & Salamoura, 2009). On the other hand, Harsch and Rupp (2011) reported the complexity of linking writing tasks to the CEFR by following the steps suggested in the Manual (i.e., familiarization, specification, standard setting, and validation) and they followed North's (2004, as cited in Harsch & Rupp, 2011 ) recommendations to align level specific writing tasks to the CEFR levels. North basically suggests that relevant CEFR descriptors in Chapter 4 (Council of Europe, 2001) that involve communicative activities are selected, relevant functional concepts and task characteristics are picked and translated into test specifications. This approach is also proposed in the CEFR: "…and in particular section 4.4 on 'Communicative Language Activities' can be consulted when drawing up a task specification for a communicative assessment" (p. 178).

The fact that the CEFR identifies language proficiency in different levels and provides illustrative descriptors of communicative activities at each proficiency level makes it a useful reference in task design especially when different levels of tasks are aimed. Even though it has also been criticized because of its limitations in language testing, it does not change the fact that the CEFR is being widely used across the world to develop comparable and standard tests as in Cambridge ESOL examinations.

#### 2.3.2.6 Test specifications

Fulcher and Davidson (2007) define test specifications as "generative explanatory documents for the creation of test tasks" (p. 52). Test specification is considered to be related with context validity since all the decisions about test content including purpose, task demands and setting are specified at this stage. It is generative because it helps item writers generate parallel items "which are similar in content and measurement characteristics" (Davidson & Lynch, 2002, p, 15); and it is explanatory because it includes all the necessary information that is needed to be able to generate equivalent test items. Regarding the usefulness and importance of test specifications, Bachman and Palmer (1996, p. 177) list four reasons: 1) they make it possible to produce parallel forms of the test with the same characteristics; 2) they offer an independent means to evaluate the intentions of test developers; 3) they allow test developers to evaluate the degree of correspondence between the actual tests and the test specifications; 4) they offer a means to evaluate the authenticity of the test. Alderson et al. (1995, as cited in Weigle, 2002) suggest that test specification differs in format and content according to the audience. To illustrate, detailed test specifications might be required for someone concerned with test validation whereas teachers may only need to have information about the general content of the test which they administer to place their students into classes.

Although various scholars have proposed several formats for specifications (Davidson & Lynch, 2002; Douglas, 2000), Bachman and Palmer (1996, pp. 172-173) suggest that test specifications for any type of test should involve: 1) the purpose of the test task; 2) the definition of the construct to be measured; 3) the characteristics of the setting of test task; 4) time allotment; 5) instructions for responding to the task; 6) characteristics of input and response, and 7) scoring method. With these key features, test specifications are used as a framework to guide multiple item writers to create a bank of equivalent items in a short time (Davidson & Lynch, 2002). Davidson and Lynch (2002) further claim the development of test specification is an iterative process, and subject to revision at any time during test development. That is to say, item writing and the process of task specification go hand in hand rather than a linear way.

In this section issues concerning context validity have been discussed. Because the idea that lies behind context validity is to prepare tasks that approximate tasks in the TLU domain, and to elicit real- life performance conditions as much as possible, it is important first to define the TLU domain to which one wishes to generalize the test tasks whether it be a university, everyday or business context. Once the types of tasks that are generally practiced in that specific TLU domain identified, decisions about task features and setting are made in the task specification process. The CEFR might be helpful in identifying the test content especially when developing tasks aimed at different proficiency levels. In the next section, the third component of Weir's validity framework, scoring validity, will be discussed.

### 2.3.3 Scoring validity

Shaw and Weir (2007) describe scoring validity as the superordinate term encompassing all the aspects of the testing procedures that are likely to influence the reliability of test scores. By this definition, it seems that scoring validity is not only concerned with the scoring process, but also with the aspects of context validity and cognitive processes along with test taker characteristics that can affect writing quality and test scores, as also discussed in the preceding sections. This shows, as Shaw and Weir (2007) note, the interconnectedness of validity components and the symbiotic relationship between these components. They further point out that scoring validity plays a critical role in test validation as tasks that are valid in terms of cognitive and contextual parameters are of little value if the marking of exam scripts is not reliable due to improper criteria or scale, inconsistent raters, lack of training, poor rating conditions etc. Shaw and Weir (2007) list the parameters of scoring validity for writing assessment drawing upon Weir's (2005) socio-cognitive framework: 1) criteria/rating scale, 2) rater characteristics, 3) rating process, 4) rating conditions, 5) rater training, 6) post -exam adjustments, 7) grading and awarding.

This section will be devoted to basic aspects of these parameters with related literature as they are considered to directly influence the reliability of scoring process.

#### 2.3.3.1 Criteria/rating scale

McNamara (1996, as cited in Weigle, 2002) states that the selection of appropriate rating criteria is of considerable importance as these criteria reflect test developer's perception of the abilities or constructs to be measured by the test. Davies et al. define rating scale as "[an assessment tool] consisting of a series of constructed levels against which a language learner's performance is judged" (1999, p. 153). The levels usually range between no mastery to full mastery of specific language skills. Each of the levels identified in the scale typically consists of verbal descriptions so that raters can determine which specific levels the written performances correspond to. A rating scale can be referred to various terms, such as assessment criteria, band level descriptors, mark schemes and scoring rubric (Shaw & Weir, 2007). Weigle (2002) argues that there are several decisions that should be made in rating scale development. In the following sections, these considerations will be addressed with reference to relevant literature.

##### 2.3.3.1.1 What type of rating scale is desired?

According to Weigle (2002) the type of rating scale to be used is the first decision that should made. Two main types of rating scales are typically distinguished in the literature: holistic and analytical (Jonsson & Svingby, 2007). These major types of scales basically have to do with the question of whether a single score or multiple scores will be assigned to each writing scripts (Weigle, 2002).

Holistic scoring involves “the assigning of a single score to a script based on the overall impression of the script” (Weigle, 2002, p.112). The major idea behind this approach is that writing is treated as a single entity in real life; therefore assigning a single score is the best way to capture the integrated qualities of writing (Knoch, 2009). Holistic scoring has been widely preferred in assessment of writing because of the advantages it offers. The rubric used for TOEFL writing assessment can be a good illustration for a holistic type of scoring rubric. Weigle (2002) discusses positive aspects of holistic scoring by citing the works of White (1984, 1985, as cited in Weigle, 2002). One of the advantages is that it is relatively faster to assign a single score with holistic scoring procedures rather than reading the scripts several times in an attempt to rate different aspects of writing. Second, test takers are not usually punished because of their poor performance on certain aspects (i.e. lexical complexity) as raters are ideally encouraged to pay attention on the strengths of the writing performance, not on the deficiencies. However, holistic scoring has also been criticized as discussed by Knoch (2009). First, it is argued that a single score is unlikely to provide diagnostic information about candidates’ ability, which is a problem especially in second language (L2) assessment. Another problem identified for L2 learners is that a single score is unable to differentiate between aspects of writing which tend to develop at different rates in L2 learners. Some candidates may be good at rhetorical aspects of writing such as development of ideas or organization while others may have mastered in linguistic skills. An impressionistic score can mask uneven aspects of L2 candidates, which might be misleading during decision making process. Furthermore, when a single score is assigned for the performances of these candidates, it is difficult to identify how raters arrive at their decisions about candidates' abilities. Sakyi (2000), for example, investigated the raters' decision making processes when evaluating writing scripts with holistic scoring procedures through verbal protocols. Sakyi identified four distinct rating styles among six raters, and it was found out that some raters did not use the scoring criteria at all and relied on personal judgments instead. While some raters focused on errors, others focused on the development of the ideas; or when they attempted to use scoring criteria, they were able to consider only one or two features to distinguish between levels of ability. The fact that raters tend to use different criteria including their personal judgments in holistic scoring may lead to inconsistencies across and within raters and finally alter the meaning of the scores (Barkaoui, 2010).

Although holistic scoring has certain advantages like being time saving and more authentic, it is obvious that using this kind of scoring procedure has resulted in concerns especially when scoring second language learners' performances. A global score assigned to a script with unbalanced qualities has the potential to punish or reward those deficient aspects. A script might be given a higher score than it deserves because of effective use of complex structures although it is quite poor in terms of content development. More importantly, various criteria raters make use of in holistic scoring may lead to inconsistent ratings, which can be accepted as a source of construct-irrelevant variance. This type of construct irrelevant variance my adversely affect score validity.

One way of overcoming possible limitations of holistic scoring is the use of analytic scoring. Analytic scoring involves assigning separate scores for various traits of writing such as content, organization, register, vocabulary, grammar; and these traits might differ based on the purpose of the assessment (Weigle, 2002). Because multiple traits are involved in analytic scoring, it is also termed as multiple-trait scoring by some researchers (i.e. Barkaoui, 2007). Analytic scoring is considered to provide more detailed information about candidates' writing abilities compared to holistic scoring. This aspect of analytic scoring is of value especially in diagnostic tests, in which detailed analysis of candidates' performance is needed for the purpose of giving feedback. Moreover, it is possible to capture uneven aspects of second language learners' writing performances through analytic scoring. Citing the works of Adams (1981) and Francis (1977), Shaw and Weir (2007) maintain that differentiation between multiple components might be particularly useful for training inexperienced raters, as analytic scoring allows raters to focus on one aspect at a time, which seems to be easier than assigning an impressionistic score. Weigle (2002) also points out that analytic scoring tends to be more reliable than holistic scoring in that multiple components of a scoring rubric are likely to increase reliability just as adding more items to discrete point test increases reliability. Because of the advantages listed above, analytic scoring is often preferred over holistic scoring by many writing specialists (Weigle, 2002). A well known example for analytic scoring is the writing band descriptors used by International English Testing System (IELTS).

On the other hand, analytic scoring is not without disadvantages. A major problem has to do with practicality. Because more than one decision needs to be made during analytic rating procedures, it takes longer time than holistic scoring. Becker (2010) investigated directors' preference of scoring rubric type in Intensive English Programs at multiple universities. The results of the questionnaire indicated that the majority of IEP directors preferred holistic scoring because of its practicality and efficiency. Secondly, Knoch (2009) maintains that there is no guarantee that raters will distinguish between multiple traits of the scoring rubric. As discussed in Myford and Wolfe (2003), it is very common that rating of one aspect may affect ratings of other aspects, creating a halo effect.

Barkaoui (2007, 2010, 2011) investigated the role of the types of rating scale (holistic and analytic) on rating processes and score variability. The findings showed that the raters using holistic scoring tended to go back to the text itself to make their decisions while raters using analytic scoring referred to the scoring rubric more frequently; in other words they were more attentive to the evaluative criteria in rating scale. Moreover, it was found out that raters were likely to be less severe with analytic marking. When it comes to reliability, holistic scoring resulted in higher inter-rater reliability whereas analytic scoring led to higher self-consistency. Barkaoui (2011) concluded that both rating types might be useful for different assessment purposes, conditions and raters.

The fact that analytic scoring can provide diagnostic information makes it valuable particularly in L2 writing assessment. Assigning distinct scores for different aspects of writing ability might be quite helpful in identifying “strengths and weaknesses of writers” (Becker, 2010, p. 116). However, due to practicality reasons, analytic scoring may not be very preferable in large scale testing, where large numbers of written responses are required to be rated in a short time. It is obvious that both holistic and analytic scoring methods have their own strengths and limitations. The testing purpose and contextual variables should determine the choice of the rating scale.

#### 2.3.3.1.2 What are the criteria based on?

When the type of rating scale to be used is decided on, the next consideration to take into account is how to design the scale itself (Weigle, 2002). The ways in which rating criteria are constructed are of great importance because the wording of the scale is considered to represent test developers’ view of writing construct to be tested, as mentioned above. Weigle (2002) discusses two methods of constructing rating scale: a priori method and empirically based methods. Weigle (2002) describes a priori method as “defining in advance the ability being measured and then describing a number of levels of attainment, from none to complete mastery” (p. 125). Knoch (2007) suggests that a priori method is carried out through intuitive judgments of experts about the nature of language development in order to construct scale descriptors. Such intuitive methods typically involve “…develop[ing] a rating scale based on pre-existing scales, teaching syllabus or a needs analysis” (Knoch, 2009, p. 43). This is generally done by a group of experienced teachers or language testers taking the role of experts. Although it is argued that most rating scales are constructed intuitively, such type of scales has been criticized by several researchers (i.e., Fulcher, 2003; Knoch, 2007; Turner & Upshur, 2002). Turner and Upshur (2002) summarize the criticisms that rating scales constructed through a priori method have received: a) the ordering of the criteria do not generally reflect the findings of second language acquisition (SLA) research; b) the criteria are mostly irrelevant to the characteristics of task response, and the context; c) the criteria are not grouped at relevant descriptor levels properly; d) the wording of scale descriptors are often ambiguous, and this causes raters to interpret the scale in a different way. These concerns raised against intuitively designed scales seem to be valuable especially for reliability of test scores and validation of rating scales. If, for example, the descriptors of a rating scale are not relevant to the task requirements and what test takers produce, raters may not be able to assign test takers' performances to any level on the scale, and refer to their personal judgment instead. In such a rating situation, the use of a rating scale would be meaningless. Ultimately, this might cause another source of construct irrelevant variance leading to score variability.

One means of responding to the criticisms of intuition-based rating scales is to construct scale descriptors empirically, which involves empirical examination of students' actual written responses and defining the characteristics that differentiate responses and the levels of rating scales (Knoch, 2007). With this method of scale development, it is claimed that the descriptors are more likely to reflect the features of test takers' performances at different proficiency levels; and describing real features of writing at each level may solve the problem of relative wording of scale descriptors. As a result, raters are expected to apply the rating scale more consistently and efficiently. What is more, empirically developed descriptors are believed to reflect the natural order of writing acquisition, and all these features are claimed to increase score reliability as raters are able to base their decisions on more explicit and realistic evidence (Knoch, 2007). Despite its promises, empirically-based rating scales have also been criticized. Fulcher, Davidson and Kemp (2011), for example, maintain that as empirically-developed rating scales involve descriptions of performance in specific genres or contexts, they may not be applicable to rate performances in other contexts. This in turn may affect the generalizability of inferences that are made based on ratings.

Knoch (2007) conducted a study with ten trained raters to investigate whether empirically developed rating scales function differently from intuition-based analytic rating scale. The results indicated that the individual components of empirically developed scale (pilot scale) were more discriminating than the intuition- based scale (existing scale). This might be a proof for the promises of empirically-based rating scales mentioned above. Moreover, the pilot scale resulted in higher inter-rater reliability whereas the raters tended to differ more in terms of severity when they used the existing rating scale. This may have to do with less explicit and relative wording of the intuitively developed scale, which caused raters to rely on their personal judgments during decision makings more often. The researcher concluded that analytic rating scales may not necessarily function in an analytic manner, if scale categories are not described explicitly and detailed enough.

When the two approaches for constructing rating scale descriptors are examined, it seems that empirically developed rating scales tend to yield more reliable rating process as they reflect the features of actual writing performance. However, the use of a priori developed rating scales is obviously more common probably because they are easier to develop and generalizable to various task types although they may function more like holistic scales. The best way in designing rating scales would be, as suggested in the CEFR, to combine all the approaches in a complementary process.

#### 2.3.3.1.3 How many points or scoring levels will be used?

Another important consideration with regard to scale construction is to decide on the number of scoring points to be used to distinguish between different ability levels. It is claimed that the number of distinctions that raters can make is limited (Weigle, 2002). Myford (2002), citing the works of several researchers (Cronbach, 1990; Guilford 1954; Linn & Gronlund, 2000; Payne, 1997) suggest that there has not been a consensus on the optimal number of scale points. In contrast to Linn and Gronlund

(2000, as cited in Myford, 2002), and Cronbach (1990, as cited in Myford, 2002) who advocated for 3 to 7 points scale, Guilford (1954, as cited in Myford, 2002) favored 11 point scales. Payne (1997, as cited in Myford, 2002), on the other hand, considered the optimal number of scale points as 7 to 9 points. This may be the reason for the fact that different numbers of scale points are employed by various large scale assessment programs. To illustrate, TOEFL IBT uses six-point scale whereas IELTS employ nine-point band descriptors.

Myford (2002) raises an important issue that raters are likely to face when using a rating scale with various levels. Myford argues that there may be some performances that fall in the cracks between levels. In other words, a written response may seem better than, for example, the characteristics of performance described at Level 1, but not good enough to fall in Level 2. For such situations, she reports the suggestions offered by Cronbach, Bradburn, Horvidtz (1994) that raters can use intermediate or midpoint values for borderline responses in order to improve the accuracy of the rating process. However, the findings of her study in which she investigated the effects of rating scales with different design features (absence or presence of midpoints) revealed no difference in student separation reliabilities.

Besides the number of scale points, it is also important to consider the number of categories if analytic scores are used. It is suggested in the CEFR that “more than 4 or 5 categories starts to cause cognitive overload and that 7 categories is psychologically an upper limit” (Council of Europe, 2001, p.123). Therefore, even though many aspects of a specific skill are defined, it may be more appropriate to select the ones that are important for the purpose and the context of the test (Weigle, 2002).

The number of scale points is an important factor to consider in scale development as the raters should be able to discriminate between test takers' various proficiency levels, and should not be overwhelmed with too many scale points, at the same time. As the literature suggests, rating scales with scale levels ranging between 4-9 and with 4 or 5 categories seem reasonable to use in order to achieve reliability. So far, issues concerning scale constructions have been discussed. Decisions regarding rating scales should be made carefully as Popham (1990, as cited in Myford & Wolfe, 2003) argued that rating scales are one potential source of error that may threaten score validity. According to Popham, another potential source of error that must be carefully monitored is the raters. Therefore, the focus of the next section will be on raters, specifically the ways that raters tend to vary in their ratings, and the factors that may cause rater variation including rater characteristics and rating process.

#### 2.3.3.2 Raters

Although Shaw and Weir (2007) used the term "rater characteristics" to examine the factors that may cause variation in the ways raters evaluate written performances, a broader term "raters" is preferred here in an attempt to investigate the relevant issues from a wider perspective.

A number of studies have shown considerable rater effects as a source of systematic variance in the ratings of written performance (i.e., Hoyt, 2000; Lumley & McNamara, 1995; Myford & Wolfe, 2003). This kind of variability is principally unwanted as "[it] is associated with characteristics of raters and not with the performance of examinees" (Eckes, 2008, p. 156). Major rater effects which are considered to be sources of systematic error variance are identified as severity/leniency, halo, central tendency, inconsistency/randomness, and bias (Knoch, 2009). These rater effects have raised concerns about validity of ratings and, thus, have been the focus of researchers (Eckes, 2005; Lumley & McNamara, 1995; Myford & Wolf, 2003; Wolfe, 2004). To illustrate, Engelhard (1994) worked with 15 raters to investigate rater effects on the quality of ratings by using many-faceted Rasch (FACETS) measurement model. The data revealed significant differences in rater severity, and two raters were found to rate the compositions holistically rather than analytically, which is the evidence of halo effect. Besides, nearly 80 % of ratings were in the two middle categories of the rating scale, displaying the presence of central tendency effect. Similarly, Eckes (2005) conducted a study to investigate rater severity and bias effects towards examinees, rating criteria and tasks in writing and speaking sections of the test of German as a Foreign Language based on many-faceted Rasch measurement. The results revealed substantial variability in raters' level of severity. Although the raters were consistent in their overall ratings, they were significantly less consistent in relation to criteria and tasks (for speaking test) than in relation to examinees. In other words, the raters were biased towards certain criteria and tasks, which led them to display more severity or leniency with those criteria and tasks.

Another line of research has investigated decision making behaviors of raters with different personal background, rating background and work experience (Knoch, 2009). Cumming (1990) examined the decision making processes of expert and novice raters and found out that expert raters used a wide range of criteria, self-control strategies and knowledge sources while reading and judging students compositions whereas novice raters tended to use much fewer of these criteria and skills probably derived from their general reading strategies, and they relied on on-line corrections of student texts to make their judgments. In a similar vein, Wolfe et al. (1998) investigated cognitive differences of proficient and non-proficient raters. The results indicated that proficient raters tended to use top-down approach through which they focused on general features of texts and made an overall judgment of writing quality. Less proficient scorers, on the other hand, seemed to use a bottom-up approach focusing on more specific features of the essay and interrupting their reading process to see if the text so far satisfies the scoring rubric.

Research also indicated variability in rating behaviors of raters with different language background and occupations. Studies investigating the effect of language background in the context of oral assessment generally revealed difference in the ratings of individual criteria although no significant difference was observed between the overall ratings of native and non-native English speaking raters (Brown, 1995; Kim, 2009; Zhang & Elder, 2014). Zhang and Elder (2014) found that native English speaking raters (NES) were more lenient on flexibility and appropriacy categories whereas non-native English speaking (NNES) raters were slightly more lenient on accuracy and range category in their judgments of speaking performance. It was concluded by the researchers that although raters from different language background tend to approach scoring criteria differently, this difference is negligible unless it affects overall ratings of candidates. In the assessment of writing, on the other hand, Johnson and Lim (2009) observed no significant pattern of language background-related bias in the ratings of MELAB writing tests, and they claimed that raters from different background can be just as effective as native –speaking raters if they are trained properly.

With regard to rater occupation, O’Loughlin (1992, as cited in Shaw & Weir, 2007) compared the rating behaviors of teachers from different subject areas who rated essays produced by native-speaker students and EFL students. Findings showed that language teachers did not pay as much attention to content as teachers of other academic subjects and EFL teachers were more attentive to grammar and cohesion than mainstream English teachers. Similarly, Weigle, Boldt and Valsecchi (2003) examined how ESL, English and other content area instructors perceive and evaluate ESL student writing. They concluded that raters from different disciplines bring their own expectations of what constitutes a good writing based on the conventions of their discourse community, which consequently influence their way of using assessment criteria. Their findings revealed, for example, that instructors of English departments were more concerned with grammar than other raters whereas psychology department raters devoted their primary focus for content.

Weigle (1999) suggests that rater expectations are another factor that may influence test scores. In a study investigating rater-prompt interactions, Hamp-Lyons and Mathias (1994) found that raters awarded higher scores to the performances in response to the tasks that were judged as difficult by the experts. Hamp-Lyons and Mathias suggested that this unexpected finding might have resulted from the compensatory strategies employed by the raters in order to negate the effect of prompt difficulty and reward students who went for the difficult tasks. Eckes (2012) examined the relationship between raters' perception of criterion importance and their rating behaviors by conducting bias analysis with multi-faceted Rasch measurement. He found that the criteria that were perceived as important received more severe ratings than the ones considered as less important.

It seems from the studies mentioned above that raters might vary in terms of their decision making processes based on their personal background, rating experience, professional training, and expectations. Research has shown that differences in rating behaviors may lead to considerable variability in scores that are not related with examinees' performance, thus threaten score validity. In an attempt to eliminate rater effects such as severity, leniency, halo, central tendency and bias, it is now obvious that one needs to construct detailed scoring criteria with unambiguous and explicit descriptors. However, a well constructed scoring rubric may not be sufficient by itself to eradicate errors associated with rater characteristics.

#### 2.3.3.3 Rater training

In an attempt to minimize variability of raters and improve the reliability of rating process, one method that has been widely practiced is rater training. Jacob et al. (1984, as cited in Weigle, 1994) argue that training "ensure more consistent interpretation and application of the criteria and standards for determining communicative effectiveness of writers" (p. 43). Shaw and Weir (2007) suggest that no mark scheme can capture the definition of a level in a way that raters could apply consistently unless each level is exemplified with benchmark scripts during rater training. Research has shown the effectiveness of rater training. Weigle (1994) investigated the effect of training on inexperienced raters of ESL compositions based on verbal protocols. The findings revealed that training helped clarify the comprehension of intended rating criteria and modify raters' expectations in terms of writer characteristics and task demands. Weigle (1998) compared the ratings of experienced and inexperienced ratings before and after training using many-facet Rasch measurement. The results demonstrated that inexperienced raters were more severe and inconsistent than experienced raters before training. Rater training proved to be successful in improving consistency of raters and reducing rater severity although significant severity was still present. In other words, rater training contributed to intra-rater reliability rather than inter-rater reliability. This finding is in line with Lumley and McNamara (1995) who found that rater training made raters more self consistent, but not eliminated rater harshness.

It is obvious that rater training is a useful means in minimizing variances in the way raters interpret and apply rating criteria which result from their personal, professional and linguistic background. Therefore, instead of being concerned about differences in rater background, it would be a wiser action to provide raters with intensive training. The fact that rater training cannot fully eradicate rater variability and improve inter-rater reliability, however, urges to conduct statistical analysis to monitor and improve rating procedures. One of these analyses that help identify sources of variability is many-facet Rasch measurement and the other is generalizability theory.

## 2.4 Many-facet Rasch measurement model

Many-facet Rasch measurement (MFRM) which is an extended version of Rasch measurement to incorporate multiple facets (variables) rather than two (traditionally examinees and items), such as raters, tasks and criteria was developed by Linacre (1989, as cited in Engelhard, 1994). It has also been termed as multi-faceted or many-faceted Rasch measurement in the literature (Eckes, 2009). It refers to “the application of a class of measurement models that aim at providing fine-grained analysis of multiple variables potentially having an impact on test or assessment outcome” (Eckes, 2009, p. 3). MFRM basically works by analyzing various facets simultaneously but independently and calibrating them on the same linear logit scale called variable map, which makes it possible to measure examinee proficiency, rater severity, prompt and criteria difficulty on the same scale and make useful comparisons. Being able to provide sample-free, scale-free measurement makes it valuable in performance assessment and rating scale validation (Schaefer, 2008). Engelhard (1992) argues that MFRM increases objectivity and fairness of measurement as it provides fair scores for each examinee that are adjusted for differences in rater severity and relative difficulty of writing tasks, resulting in a more accurate picture of writing ability. Another feature of MFRM is that along with obtaining group level main effects of facets; MFRM approach allows researchers to investigate individual- level effects of elements within a facet. For example, MFRM analyses not only provide information about raters' differences in severity (group-level), but it also pinpoints which particular raters display severity or leniency towards which particular criteria (Myford &Wolfe, 2003).

MFRM also calculates fit statistics (infit and outfit) that indicate how well the values of individual examinee, rater and task match with the expected values that are estimated by the model itself (Sudweeks, Reeve & Bradshaw, 2005). Infit and outfit mean square values have an expected value of 1 and range from 0 to infinity. According to White and Linacre (1994, as cited in Sudweeks et al., 2005), a reasonable range of infit and outfit values is between 0.6 and 1.4; values between 0.5 and 1.5 are considered productive for measurement values; and between 1.5 and 2 are considered not productive but not degrading whereas values above 2 are distorting. Engelhard (1992) suggested values less than 2 are acceptable fit values.

Finally, MFRM allows for bias analysis, which investigates interaction patterns between certain facets. To illustrate, it shows whether one rater tended to assign scores differently on a certain criterion than others or whether a group of examinees performed differently on a certain task. Each interaction between facets is assigned a bias score on a logit scale, the significance of which is displayed with a z score. A z score equal to or greater than 2.0 indicates significant interaction. (Sudweeks et al., 2005). Schaefer (2008) claims that understanding source of rater bias through bias analysis can help improve rater training and rating scale development.

## 2.5 Generalizability theory

Generalizability theory (G theory) is another measurement model that can estimate multiple sources of error in test scores. Webb and Shavelson (2005) define G theory as “a statistical theory for evaluating dependability (or reliability) of behavioral measurements” (p. 599). Dependability here has to do with the accuracy of generalizations made from a person’s observed score to the score he or she would receive under all possible conditions ( in the universe of admissible observations): a universe score, which corresponds to true score in classical test theory (CTT) (Webb & Shavelson, 2005). G theory was developed by Cronbach et al. (1972, as cited in Schonen, 2005) as an extension of classical test theory.  As opposed to CTT, which decomposes observed score into a true score variance (systematic) and undifferentiated error variance (random), G theory is able to disentangle multiple sources of variance and their interactions in scores through the application of analysis of variance (ANOVA). That is why Brennan (2000) calls CTT and ANOVA as parents of G theory.

In the case of writing, G theory could be used to decompose total variability in the scores into variance components (facets) based on sources of variation including 1) systematic variability between individual writers; 2) variability between raters (inter-rater inconsistencies); 3) variability within raters across rating conditions (intra-rater inconsistencies); 4) variability between tasks (Sudweeks et al., 2005, p.

241). Along with providing main effects of variance components, G theory can estimate interactions among the variance components (i.e., person x task, person x rater, person x task x rater). Obtaining information about relative contribution of each variance components to measurement error enables researchers to pinpoint which sources of variance is the most problematic and need to be addressed (Sudweeks et al., 2005).

G theory distinguishes two types of decision-making situations (relative and absolute decision) and produces statistics accordingly. It provides a number of statistics such as relative error variance, absolute error variance, generalizability coefficient (g- coefficient) for relative decisions, and dependability coefficient (the phi-coefficient) for absolute decisions. The g-coefficient and the phi-coefficient are analogous to reliability coefficients in CTT and range between 0 and 1 (Sudweeks et al., 2005). The g coefficients are useful in making decisions about examinees based on their relative rank ordering compared to others whereas phi coefficients are used to decide whether examinees have achieved a predetermined level of performance (Lee & Kantor, 2005).

In the case of multiple raters and tasks employed in the assessment of writing proficiency, a fully crossed, two - facet design (p x t x r) is considered to be the most powerful G study design. In such a design, all the examinees are required to take all the tasks, which are in turn rated by all the raters. However, as Lee and Kantor (2005) claim this type of design may not be feasible for scoring a large number of examinee responses particularly in large scale tests. For this type of situation they suggest an alternative design in which ratings (r’) is considered as a random facet instead of raters (r). As all the examinees’ responses are assigned two different ratings, it could be possible to use a fully crossed design even if not all the raters rate all the responses of all examinees. In addition to G studies, G theory allows for D studies through which the effects of different combinations of facets on the score dependability can be investigated (Brindley, 2000). D studies enables researchers to estimate the optimal number of tasks, raters or items to obtain desirable level of score dependability.

Although the two measurement models (MFRM and G theory) are used to estimate sources of variances involved in a test score, they do not provide the researchers with the same kind of information. G theory, for example, does not adjust scores for task difficulty and rater severity as MFRM analyses do. Second, whereas G theory only gives the main effects of different variance components in score variability, it is possible to obtain individual-level information and pinpoint specific cases through many facet Rasch analyses. Besides, the MFRM assigns an ability estimate for each student called a measure which has the features of an equal interval scale, and differentiate it from the original raw scores (ratings), which are typically ordinal (Sudweeks et al., 2005). This feature of the MFRM is considered to be one advantage, which is not achieved by the G theory as the G theory conducts analysis assuming that the data have interval scale properties (Sudweeks et al. 2005). Therefore, several researchers have used G theory in combination with MFRM to obtain detailed information from different perspectives to examine score validity (Brindley, 2000; Harsch & Rupp, 2011; Sudweeks et al., 2005).

## 2.6 Conclusion

Validity and reliability are considered to be two major qualities in evaluating usefulness of a language test. Weir (2005) provides a comprehensive framework that enables test developers to collect validity evidence during test development (a priori evidence) and after test administration (a posteriori evidence). It is argued that three components of this framework (i.e., cognitive validity, context validity and scoring validity) are critical in any language test to provide theoretical and empirical evidence for the appropriateness and usefulness of the tests (Shaw & Weir, 2007). Therefore, the present study sets forth to provide evidence for the validity of the newly developed TFL writing test in relation to these three components of validity.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

In this chapter, the methods and procedures that have been followed both during the test development and the main study will be reported. The chapter will present information about the aim of the study, instruments, data collection procedures and rating procedures. Finally, the chapter will close with the research questions and the data analysis that were used to investigate the questions.

## 3.2 The aim of the study

The aim of the current study is to investigate test usefulness of a newly developed writing test for learners of TFL as suggested in the model proposed by Bachman and Palmer (1996). To this end, a comprehensive validity framework offered by Weir (2005) has been utilized to provide theoretical and empirical evidence for test validity as validity is considered as an essential indication of test usefulness. However, as discussed in chapter 2, test validation is a comprehensive and ongoing process involving various aspects and stages before and after the test event. Although it does not seem to be feasible to address all aspects of validity in one study, attempts will be made to provide evidence for scoring validity as well as provide arguments for cognitive and context validity through qualitative and quantitative analyses for the TFL writing test in this study.

Although the study mainly focuses on the a posteriori validation of the newly developed TFL test of writing, an important part of the study consists of the initial development stage of the test which can be considered as its pilot phase. The pilot phase will be briefly summarized here to give the reader information on the rationale and procedures underlying the test in addition to the cognitive and contextual validity arguments in the following chapter. However, as the main study is going to focus mainly on scoring validity, the results from the different stages of the pilot phase will not be reported in the results section. Below is the summary of the procedures that were followed during the pilot phase briefly given under two titles: Task development and the rating scale development. The chapter then presents participants, data collection and scoring procedures and the research questions concerning the main study that will be reported in detail in this study.

### 3.2.1 Task development

The development of writing tasks was an iterative process including stages such as identifying task specifications, producing tasks, consulting experts, pilot tests and revising.

Initially, the TFL writing test was intended to be an academic proficiency test geared at B1, B2 and C1 levels as defined in CEFR (Council of Europe, 2001, p. 23). This was achieved by selecting task types in accordance with "Can Do Statement" that are specifically described for writing domain for each proficiency level, pp. 71-72). Table 1 presents the CEFR descriptors for B1, B2 and C1 levels writing skill.

Table 1. The CEFR Level Descriptors for Writing

| CEFR level | Descriptors |
|---|---|
| B1 | Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest.<br>Can write accounts of experiences, describing feelings and reactions in simple connected text.<br>Can write a description of an event, a recent trip – real or imagined.<br>Can narrate a story.<br>Can write short, simple essays on topics of interest.<br>Can summarize, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his/her field with some confidence |
| B2 | Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources.<br>Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned.<br>Can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail.<br>Can evaluate different ideas or solutions to a problem. |
| C1 | Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.<br>Can write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. |

Source: Council of Europe, 2001 pp.71-72

These descriptors were evaluated by the TFL course instructors to identify target language use domain thus the researcher was able to design tasks which were similar to the tasks used in the class. In other words, the course instructors chose the statements that they thought relevant to their students. With the information gathered from the course instructors, through the analysis of course materials and CEFR descriptors, test specifications were written (See Appendix A). Next, two or three tasks for each proficiency level were generated and presented to an expert who specializes on teaching TFL for feedback. Based on the feedback, necessary changes on the task type, instructions and prompts were made. The first draft of the test consisted of 7 tasks as summarized in Table 2. All the drafts of writing tasks are presented in Appendix B.

Table 2. The Summary of Initial Tasks

| Tasks | Targeted CEFR Level | Genre |
|---|---|---|
| 1 | B1 | Writing an e-mail to describe a place |
| 2 | B1 | Writing a story |
| 3 | B2 | Writing a report (graph interpretation) |
| 4 | B2 | Writing a film review |
| 5 | B2 | Writing a problem & solution essay |
| 6 | C1 | Writing an argumentative essay |
| 7 | C1 | Writing an argumentative essay |

These tasks were pretested on a small group of test takers to identify which tasks functioned well or poorly so that they could be revised accordingly. It was planned to obtain at least 5 samples of each task from students in TKF 212 (B2) course in order to diagnose any possible problems with the tasks. This group of students was considered to be suitable for this purpose as they would be able to deal with the tasks from the upper level (C1) and the lower level (B1). However, because it was about the end of school term in May 2014, very few students showed up on the day of test administration, which led to fewer samples than initially planned. Table 3 illustrates the number of responses for each task.

Table 3. Number of Responses Obtained

| Tasks | Task level | Genre |
|---|---|---|
| 1 | B1 | 1 |
| 2 | B1 | 3 |
| 3 | B2 | 3 |
| 4 | B2 | 2 |
| 5 | B2 | 3 |
| 6 | C1 | 1 |
| 7 | C1 | 1 |
| Total | | 14 |

Along with the tasks, the students also completed a task evaluation form in order to give feedback on task quality. The questions in the task evaluation form were aimed to explore clarity of instructions and prompts, task difficulty, adequacy of allotted time and expected text length, and the relevance of tasks and topics to test takers' academic life and specifically to their Turkish courses. The task evaluation form is provided in Appendix D.

Based on the information and feedback obtained from the first pilot testing, Task 1, Task 3 and Task 7 were kept and revised for the second pilot testing, and the other tasks were abandoned. Therefore, the second draft of the test consisted of 3 tasks as summarized in Table 4.

Table 4. The Summary of Tasks in Second Draft

| Tasks | Targeted CEFR Level | Genre |
|---|---|---|
| 1 | B1 | Writing a story |
| 2 | B2 | Writing a report (graph interpretation) |
| 3 | C1 | Writing an argumentative essay |

A second pilot testing was administered with the aim of obtaining further information on the task quality including prompts, instructions, genre, allotted time period, and text length in July 2014. Twenty four students took part in the second pilot testing. The students were TFL for academic purposes students at Summer School at Boğaziçi University at the time of testing. The tasks were administered by their course instructors in different sections as a part of course requirement rather than as an actual proficiency test in which test takers would take all the tasks in the given time. Moreover, there were not any time constrains during the administrations. Instead, students were asked the write down the duration they needed to complete the tasks in order to identify the optimal time period to complete the tasks. The second pilot test was valuable in that all the test takers answered all the tasks though in different sections, and the test takers' Turkish proficiency included C1 level besides B2, which gave rise to a variety of performances in the data. The responses from all the test takers on all the tasks along with their feedback constituted important data for task and rating scale quality. In the light of students' verbal and written feedback during the exam, the course instructor's observations, and the analysis of students' responses, the graphs in Task 2 (writing a report-graph interpretation) were completely changed, and the instructions were improved as the instructions of Task 2 turned out to be ambiguous and led to different response formats. The instructions of Task 3 were also improved to reflect the argumentative nature of the expected response. Task 1, on the other hand, was completely abandoned based on the feedback obtained through expert opinion form, which was administered to two experts specializing in applied linguistics and assessment after the second pilot testing to obtain feedback on the test content, instructions, task difficulty and scoring criteria (See Appendix E). The experts considered that writing a story (Task 1) is more suitable to assess students' general writing skills rather than academic skills. As a result, only two tasks were retained for the main study as presented in Table 5.

Table 5. The Summary of Tasks in Phase 3

| Tasks | Targeted CEFR Level | Genre |
|---|---|---|
| 1 | B2 | Writing a report (graph interpretation) |
| 2 | C1 | Writing an argumentative essay |

The time and the expected text length were finally set and the test format was finalized for the two tasks that were kept for the main study (the graph interpretation task and the essay task).

### 3.2.2 Rating scale development

Just as the task development, the development of the rating scale was iterative consisting of several stages such as trialing, revising, and multiple drafting before its final version. It was developed as an analytical rubric with four assessment criteria: content, organization, language use and vocabulary. Each criterion has four levels and each level has two categories, so the score that can be assigned for each criterion ranges from 1 to 8.

The rating scale was initially constructed based on the adaptations from the ESL Composition Profile proposed by Jacobs et al. (1981, as cited in Weigle, 2002), IELTS band descriptors for Task 1 and Task 2 and Written Assessment Criteria Grid proposed by Council of Europe (2009). Weigle (2002) maintains that Jacobs et al.'s (1981) ESL composition profile has been one of the most influential and most widely employed analytic scales. Several researchers made use of this profile in their study directly or by adapting it to rate the writing tasks in their study (i.e., Bacha, 2001; Delaney, 2008; East, 2009; Ong & Zhang, 2010). This is basically the reason that it was chosen by the researcher for the present study. IELTS band descriptors for Task 1 and Task 2 were considered relevant to the content of the rating scale developed for this study as the purpose of IELTS writing component is to assess test takers' academic writing abilities by means of the tasks that are similar to the tasks used in the current study (A graph interpretation task and an independent essay task). The descriptors of Written Assessment Criteria Grid were also analyzed as the grid analytically portrays the expected performance descriptors for each of the six proficiency levels in the CEFR from several aspects (i.e., overall, range , coherence, accuracy, description, arguments). Although it was not the primary concern of the researcher to align the rating scale to the CEFR proficiency levels, the descriptors especially for range, accuracy and coherence were useful in setting the criteria for level specific differences (i.e., to determine what features of student performance make excellent to very good level different from good to average level for the language use component).

However, after the first trial, it was observed that the descriptors of the initial draft scale were not good enough to capture the features of student responses from various task types. Regarding this issue, one assessment expert, who also rated student responses using the initial draft of the scale commented that the task requirements may change for different task types, and thus, they should be made clear in the scale. After this important feedback, the initial draft scale went through considerable revision through the analysis of student responses from the first and the second pilot testing under the consultation of the expert. Through the analyses of student responses, certain features that correspond to each scale level for each criterion were identified. The wording of the descriptors was made as explicit as possible in order not to lead any ambiguity.

The final draft consisted of two distinct rating scales to be used for the graph interpretation task and the argumentative essay task in an attempt to reflect the relevant constructs that each task was intended to represent. All the drafts of the rating scale is presented in Appendix C. Therefore, the final draft of the rating scales was based on empirically-based approach as well as a priori approach. In her study, Knoch (2007) found out that the descriptors of empirically-developed scale were more discriminating and resulted in higher inter-rater reliability and self-consistency among the raters. Therefore, it was expected that the updated scales would contribute better to score validity.

All the procedures and stages reported above led the researcher to obtain the final version of the rating scales and the final draft of the two writing tasks to be used in the main study. The methodology for the main study will be reported below.

## 3.3 Participants

Forty seven students who came to Turkey for one or two semesters through International Student Exchange Programs from different countries participated in the main study. The students were registered in Turkish for Foreigner (TKF) classes offered by Turkish Language and Literature Department at Boğaziçi University. There are several TKF courses for students with different proficiency levels in Turkish. TKF 111 and 112 courses are offered to students with low levels of Turkish proficiency corresponding to A1 and A2, TKF 211 and 212 are available to students with intermediate levels of Turkish proficiency (B1 & B2), and TKF 315 and 317 courses are for students at advanced levels of Turkish proficiency. However, students are placed in these classes not by a standardized placement test, but by students' own perception of proficiency in Turkish or by the judgments of instructors teaching TKF courses. In this study, only students enrolled in TKF 211, 315 and 317 classes took the exam as the exam is aimed for intermediate and advanced levels of proficiency in Turkish. Just before the examination started, the participants were asked to fill out a participant profile form to obtain information on their age, country of birth, native language, language of education, length of learning Turkish, length of stay in Turkey and their level of Turkish proficiency in five skills (reading, listening, writing and interaction). The participant profile form is provided in Appendix F. Two of the participants did not fill out the learner profile information, so the information is presented for 45 participants. Of all the participants, 19 were male and 26 were female. Their age ranged between 21 and 33 with a mean of 23. In terms of their country of birth and language background the participants constituted a diverse group. They came from 19 different countries with the 4 largest groups being from Japan (20 %), Germany (13.3%), Greece (8.9%) and the USA (8.9%). There were 10 participants who were exposed to Turkish at birth or at an early age although they were born and grew up in countries other than Turkey. Most participants had been learning Turkish for more than 1 year by the time the examination took place. The average length of their learning Turkish was 75 months, however if those ten participants who had been learning Turkish since birth or childhood were excluded, the average decreased to 24 months. Their average length of stay in Turkey was 17 months, which means that most of them started learning Turkish before they came to Turkey.

In the last part of the participant profile form, participants were asked to evaluate their proficiency level in writing through CEFR Self-Assessment Grid (Council of Europe, 2001, pp. 26-27). Accordingly, the average writing proficiency level was 2.4 for 14 participants from TKF 211 class, 3.4 for 23 participants from TKF315 class and 4.1 for 7 participants from TKF 317 class, which correspond to A2, B1 and B2 levels on CEFR scales respectively.

## 3.4 Data collection procedures

The data that constituted the basis of the present study was collected through the main administration that took place in participants’ usual classes in the arranged time in October 2014. Participants were informed about the test by their instructors before the administration. Two researchers, who were working as research assistance at Boğaziçi University at the time, were responsible for the administration. They were well informed about the test and the procedures. Because one class hour was granted to administer the test in each class by the course instructors, participants had 50 minutes to complete the two writing tasks (20 minutes for task 1 and 30 minutes for task 2). They were also required to fill out the task evaluation form after completing each task.

### 3.5 Scoring procedures

The writing scripts obtained from the main study were marked by two native speakers of Turkish and the researcher. Two of the raters involved in the scoring were working as research assistant in English Language Teaching department at Boğaziçi University and did not have much experience in rating writing scripts by using a rating scale when the rating procedures took place. The third rater was the researcher herself. She worked as an English instructor at a private university and taught English to new coming students in the preparatory year. Compared to the other two raters, the researcher was more experienced in rating as her job involved rating students' writing responses obtained from of a variety of tasks at different proficiency levels. She was also familiar with different types of rating scales (i.e., analytic vs. holistic) to rate student responses as part of achievement tests, proficiency tests and writing portfolio assessment. However, none of the raters were experienced in rating responses that are written in Turkish by L2 learners of Turkish. In this aspect, the rating of student responses was a new experience for all the raters, and the rating scales developed for the study were also new for the two inexperienced raters. Therefore, an intensive rater training session was conducted before the actual rating session.

The aim of the rating training was to familiarize the raters with the rating criteria and help them interpret and apply the rating scale in a consistent way. In other words, rating training was conducted in an attempt to minimize rater-related variances, and improve the reliability of scoring procedures. During the training session, the raters were first informed about task demands, writing construct and rating procedures. They were then familiarized with the scale descriptors and encouraged to ask any questions regarding the wording of the descriptors. A bunch of benchmark scripts that represent different scale levels for each task which had been previously chosen and rated by the assessment expert and the researcher were assigned to the raters for rating. Once they finished, they compared their scores with the scores given by the researcher and the expert. With a thorough discussion of why the raters gave those scores, which features of the responses they paid attention to while assigning scores for each of the individual criterion in the scale, it was aimed to arrive at a mutual understanding of how to use the rating scale during rating.

After the training session was completed, each script was double-scored by two raters to ensure inter-rater reliability. The two scores assigned to each script were then compared for discrepancy. When significant differences were observed between the first and the second rater, they were asked to score the scripts for the second time without seeing the initial scores. With this way, the researcher aimed to minimize the discrepancies between the ratings.

## 3.6 Research questions and data analysis

As noted before, the study aims at validating the newly developed Turkish test of writing for TFL students by attempting to provide evidence for cognitive, context and scoring validity aspects.

In order to establish the differing cognitive demands placed by the two tasks and contextual features of the tasks, the following question has been formulated.

Research question 1. How do the graph interpretation and essay tasks differ from each other in terms of cognitive and linguistic demands?

In order to address this research question, four sub-questions have been raised:

1.1 What are the qualitative differences between the tasks in terms of cognitive and contextual features?

The question is concerned with both a priori (cognitive and context validity) and a posteriori validation. It is hypothesized that the two types of academic writing tasks differ from each other in terms of their cognitive and linguistic demands. An extended description and justification of the cognitive and contextual characteristics of each task provided evidence for cognitive and contextual differences between the tasks.

1.2 Do the students perform differently on these task types?

For the investigation of this question, the mean scores from the two tasks were compared through a paired samples t-test. The effect size was calculated using Cohen’s *d* (Pallant, 2007). It is hypothesized that there will be a statistically significant difference between students’ mean scores on the two tasks and the students will score higher on the B2 level graph interpretation task in comparison to C1 level essay task.

1.3 Are the tasks different in terms of difficulty?

In order to investigate this question, many-facet Rasch measurement (MFRM) analysis was implemented using Minifac (FACETS) software, version 3.71.4 which was developed by Linacre (2014). As this method of analysis is relevant to several analyses used in response to further research questions as well, it is appropriate to introduce MFRM model and the indices that were used:

Many-facet Rasch analysis: The MFRM analysis provides a variable map in which students, raters, tasks and criteria are calibrated on the same logit-scale with equal intervals. The variable map gives information about student distribution based on their proficiency, rater severity, and task and criteria difficulty. Along with the variable map, the MFRM analysis produces a measurement report (i.e., rater measurement report, examinee measurement report) for each facet involved in the analysis (Eckes, 2009). These measurement reports include significant statistics such as fit indices (infit and outfit mean square values), fixed effect chi-square tests, and two different separation statistics: The separation index and the reliability of separation index (Myford & Wolfe, 2003).

Fit indices show the extent to which the observed measures of students, raters and tasks match with the expected measures that are estimated by the many-facet Rasch measurement model (Myford & Wolfe, 2003). Fit indices consist of infit and outfit mean square values. Possible mean square values for infit and outfit indices range between 0 and 1. Linacre (2008, as cited in Eckes, 2009) suggests that values between 0.5 and 1.5 are "…productive for measurement or …indicative of useful fit" (p. 18).

Fixed effect chi-square test indicates "[whether] the fixed effect hypothesis that the estimates of all the elements within a given facet can be viewed as sharing a common parameter, after allowing for measurement error" is true (Myford & Wolfe, 2003, p. 409). For example, the fixed effect chi-square test for tasks tests the hypothesis that all the tasks in the study are of equal difficulty.[1] The significance value reported shows the probability of whether the fixed effect hypothesis should be kept or rejected.

The separation statistics reports "the amount of variability (or spread) in the measures estimated by the MFRM model for the various elements in the specified facet relative to the precision by which those measures are estimated" (Sudweeks et al., 2005, p. 245). The reliability of separation index can range between 0 and 1, whereas the value of the separation index ranges between 1.0 to infinity. These two statistics are reported for each facet by the MFRM, and they are interpreted differently for each facet (Sudweeks et al., 2005).

In order to investigate the research question 1.3, relative difficulty estimates of tasks from MFRM analysis were used. Specifically, task difficulty statistics from the variable map, the reliability of task separation index and fixed chi-square test for tasks from the task measurement report were employed. The more difficult task appears higher in the column while the easier task appears lower in the variable map. A value that is close to 1.0 for the reliability of task separation index is an indication that the tasks differ from each other in terms difficulty, which can be further confirmed if the test of fixed chi-square value is significant and the fixed effect hypothesis (all the tasks are of equal difficulty) can be rejected (Sudweeks et al. , 2005).

In the present study, it is hypothesized that the tasks will be different in terms of difficulty and the essay task will be more difficult. Therefore, it is expected that

[1] Similarly, 'the fixed effect chi-square test for the raters facet tests the hypothesis that all raters exercised the same level of severity when evaluating ratees, after accounting for measurement error'(Myford & Wolfe, 2003, p. 409) .

the essay task will appear higher in the map and will have a higher value of difficulty measure. The reliability of separation index will be close to 1.0 and the chi-square test will yield significant result.

1.4 Do the tasks reliably separate students into distinct proficiency categories? To examine this question, student distribution of writing proficiency estimates from MFRM analysis were employed. Infit and outfit mean square values of students, student separation index, the reliability of student separation index and the test of fixed chi-square were employed to examine how well the students were discriminated based on their proficiency by the two tasks. Fit indices for students can be used to identify unusual ratings assigned to student responses (Park, 2004). Infit and outfit mean square values that are under 0.5 or over 1.5 are accepted as misfitting students who display unusual rating profiles.

In the present study, it is hypothesized that the tasks will reliably separate students into distinct proficiency categories. Therefore, a high value of student separation index and a value of reliability separation index close to 1.0 are expected to claim that the tasks reliably separated the students into different levels of proficiency.

In order to provide evidence for scoring validity of the test, three questions have been formulated. As the rating scales used to assess students’ responses have the potential to affect the reliability of test scores, the second research question investigates the reliability of the rating scales.

Research question 2. How reliably does the rating scale function for this specific group of raters?

In order to obtain reliable scores, it is hypothesized that the rating scales will function reliably for the group of raters involved in the ratings. In other words, the rating scales will lead to consistent ratings within and among the raters.

For the investigation of the question, selected statistics from criterion measurement report, inter-rater and intra-rater reliability statistics, and category statistics that were produced by the MFRM analysis were used:

Selected statistics from the criteria measurement report: Statistics of criteria measures, the criteria separation index, the reliability of criteria separation index and fit indices for criterion were reported. Criteria measures (in logits) give the same information with the variable map, which includes a column on criteria difficulty. On the variable map, the scale criterion[2] that appears higher in the column indicates the hardest criterion for students to receive high scores on, and the criterion that appears lower on the column shows the easiest criterion to receive high scores on. Similarly, the higher criterion measures show the more difficult criteria for students to get high scores. The criteria separation index is used to identify the number of statistically different strata of criteria difficulty, which might be used to determine if the raters actually apply the rating scales in an analytical way or not (Myford & Wolfe, 2004). The reliability of separation index for criteria is expected to be closer to 1.0 as the criteria in a scale are supposed to be of differing difficulty as an indication of analytical functioning. The infit and outfit mean square values are expected to be close to 1.0 between the range of 0.5 and 1.5 in order to argue that they all relate to the same construct (unidimensionality) (Eckes, 2009).

[2] In this case, the criteria in the scale are content, organization, language and vocabulary.

Inter-rater & intra-rater reliability estimates: For inter-rater reliability, the point biserial correlation indices were used, and rater fit statistics were examined to find evidence for intra-rater reliability. The point biserial correlation measures need to be close to 1.0 for high inter-rater reliability (Knoch, 2007). Rater infit and outfit mean square values are expected to be between 0.5 and 1.5 in order to claim that raters used the rating scale consistently (intra-rater reliability) (Eckes, 2009)

Selected category statistics: The MFRM provides several statistics in order to evaluate the effectiveness of the rating scale. The average student measures, outfit mean-square values and Rasch Andrich Threshold measures were used to examine scale effectiveness (Eckes, 2009). The average student measures are required to advance monotonically as the scale categories increase from 1 to 8 in order to claim that the scale categories are appropriately ordered and meaningfully applied. Similarly, Rasch Andrich Threshold measures need to increase with the category scales. Outfit mean square values are supposed to be smaller than 2.0 to argue that the categories are used appropriately by the raters.

Raters are known as another potential source of variability that lead to construct irrelevant variance and lower score validity. The third question, therefore, is concerned with the potential involvement of rater effects in the ratings of student responses.

Research question 3. To what extent is the quality of ratings influenced by rater effects?

In order to provide evidence for scoring validity, it is hypothesized that no significant rater effect will be involved in the ratings of student responses. The rater effects were investigated in terms of rater severity, halo effect, central tendency and inconsistency. The following statistics from the MFRM analysis were used for data analysis:

Rater severity: To examine if there was a considerable difference in the severity of raters, rater severity measures, rater separation index and the reliability of rater separation index from the rater measurement report were used. Raters with higher measures (in logits) appear to have exercised higher levels of severity than the ones with lower severity measures. To be able to claim that raters involved in the ratings exercised similar levels of severity, the difference between the most severe and the least severe rater should be as small as possible. In addition, rater separation index is expected to be close to 1.0 to argue that the raters were interchangeable as the index shows the number of statistically distinct groups in terms of severity (Eckes, 2009). Finally, the reliability of rater separation index need to be close to 0 as this index indicates how separate the raters are in terms of severity they exercised (Myford &Wolfe, 2003). The closer the reliability index to 1.0, the more different the raters are in terms of severity.

Inconsistency: To investigate inconsistency that the raters might have displayed in their ratings, rater fit indices were reported. If infit and outfit mean square values are within the range of 0.5 and 1.5, one can argue that the raters were self consistent in their ratings.

Central tendency: Rater fit indices and student separation index were used to investigate central tendency. When the infit and outfit mean square values are lower than 0.5 (overfitting raters), one can conclude that raters tended to overuse certain categories, which is accepted as an evidence for central tendency when raters are generally thought to overuse the middle categories (Knoch, 2007; Myford & Wolfe, 2004). Student separation index is another statistics that can show the existence or absence of central tendency. A high student separation index means that students were well discriminated in terms of their levels of proficiency by the raters, and thus it can be used as another evidence for the absence of central tendency (Sudweeks et al., 2005).

Halo effect: To examine halo effect, the criteria separation index from the criteria measurement report was used. In order to argue that raters were able to distinguish between conceptually different aspects of rating scale (scale criteria), the criteria separation index is expected to correspond to the number of criteria in the rating scale.

In order provide additional evidence for score validity, reliability of test scores were investigated in research question 4 through the analysis of generalizability theory.

Research question 4. How dependable are the scores assigned to the examinees?

4.1 What is the relative contribution of persons, tasks and raters to the overall variability in the ratings?

4.2 How many writing tasks and raters are necessary to achieve /increase acceptable levels of reliability?

Generalizability theory analysis makes it possible to estimate the magnitude of variability resulting from tasks and raters. In order to claim the reliability of the test scores, it is expected that most of the variance will result from students, as they are naturally expected to differ in their abilities.

Generalizability theory analysis: G theory analysis consisted of two stages that include generalizability and decision studies (G-studies and D-studies). EduG-6e software program was employed to conduct G and D studies. In order to estimate main and interaction effects of variance components on the observed score variance, a two- facet crossed design (p x t x r') in which persons crossed with tasks crossed with ratings, with tasks (t) and ratings (r') as random facets was used. The design (p x t x r') featured a fully crossed univariate design and it was conducted for each of the four criteria. In the G-studies, ratings (r') - instead of raters (r) which is commonly used in G studies - were preferred as a random facet. As Lee and Kantor (2005) suggested and applied in their own study, ratings as a facet make it possible to use a fully-crossed design in conditions in which it is not possible for all the raters to mark all the responses although all the responses are assigned two ratings. D-studies were conducted to investigate the effects of different numbers of task and rating scenarios on score dependability[3]. Changes in the G coefficients were reported to examine the optimal numbers of tasks and ratings in order to maximize the dependability of the analytic scores.

## 3.7 Conclusion

In this chapter, the procedures and methods that were followed during the main study were reported in detail after a brief summary of the procedures that were followed in the initial development stage of the test (pilot phase). The research questions of the study were, then, presented with the relevant statistics that were used to investigate each question. In the following chapter, the results from the analyses that were used to investigate the research questions will be presented.

[3] For example, how is reliability affected when the number of tasks or the number of ratings is increased?

# CHAPTER 4

# RESULTS

## 4.1 Introduction

As discussed in the previous chapters, the study was conducted in order to provide evidence for the validity of newly developed TFL writing test to support its usefulness. With this aim, four questions were formulated to investigate both a priori and a posteriori aspects of validity in terms of cognitive, context and scoring validity. In this chapter, these four questions will be investigated by presenting the results from the relevant analysis of the data.

## 4.2 Research question 1: How do the graph interpretation and essay tasks differ from each other in terms of cognitive and linguistic demands?

### 4.2.1 Cognitive and contextual task features

Qualitative analysis reported here includes the cognitive and contextual characteristics of the graph interpretation and essay task as operationalized by the researcher. These two different task types are intended to elicit responses that reflect the kind of writing skills that are required in academic contexts. The graph interpretation task is considered to be an integrated task and the essay task is thought to be a stand- alone (independent) task. The difficulty levels of these tasks were assumed to correspond to B2 and C1 proficiency levels as proposed in the CEFR. In other words, these two different task types were not intended to be comparable in terms of difficulty. This was achieved by operationalizing differing cognitive and contextual demands and features for each task. Such differences in cognitive and contextual characteristics are assumed to give rise to differences in test takers' performance.

### 4.2.1.1 The graph interpretation task

The graph interpretation task requires students to write a short description of information provided in the form of a bar graph. The visual data consist of two graphs so that students can synthesize information from the two sources and make the necessary comparisons. The graphs were chosen among a bunch of alternatives after trying each of them out with several native and non-native speakers of Turkish. The first graph presents the data about people holding doctoral degrees in Turkey in 2008 based on their ages and genders. The same information is given in the second graph based on the disciplines and genders of the people in the data (See Appendix B). The main trends in the data are clear and distinguishable, and the topic is relevant to students' lives. It was among the main concerns of the researcher not to privilege any group of students due to topical knowledge. Since the students participated in the study were all university students, a task about people holding doctoral degrees is expected to appeal to their interest. The main sub-skills required to complete the graph interpretation task are assumed to be: 1) identifying the main trends in the data; 2) selecting important information; 3) making comparisons; 4) organizing the information in an order of significance. These requirements of the task were made clear to the students in the instructions by asking them to describe important information in the graph and make comparisons when necessary in their reports. The cognitive demands specifically involved in this task are graph comprehension, graph interpretation and graph translation as suggested by Yang (2012) in the study in which Yang examined test takers' academic writing strategies that were employed during the completion of a graph-writing test.  In other words, students needed to comprehend the visual data presented in the form of a graph, identify key points and translate them into written piece of text through planning, evaluating, monitoring and revising in order to complete the task. That is why, it is considered to be an integrated task.

With regard to response characteristics of the task as one aspect of contextual features, students were expected to write at least 150 words in 20 minutes. In terms of linguistic requirements, the graph interpretation task is associated with task-related vocabulary and relatively limited range of grammatical structures and language functions related with data presentation, source identification, and comparison. Table 6 presents some examples of vocabulary and grammatical structures from the students' responses.

One can see from the extracts from student responses that certain lexical items were used by various students to write their reports. Similarly, students frequently used present simple tense, comparative/superlative forms, and less frequently relative clauses, noun clauses, and if clauses in their responses to fulfill certain language functions required by the tasks such as data presentation, source identification and making comparison.

Table 6. Samples from Student Responses on the Graph Interpretation Task

| Extracts from student responses | English equivalents | Grammatical structure |
|---|---|---|
| *…ikinci grafiğe göre, en yüksek oran Tıp ve Sağlık bilimleri dalından.* | …according to the graph, medical sciences has the highest percentage. | Present Simple |
| *…doktora dereceli erkek ve kadınların istatistiklerini gösterir.* | It shows statistics of men and women with doctoral degrees. | Present simple |
| *Bu bireylerin çoğunluğu erkeklerden oluşmaktadır.* | Most of these individuals consist of men | Present simple |
| *İleri yaşlarda doktora dereceli bireyler daha azdır* | There are fewer individuals with doctoral degrees at older ages. | Present simple, comparative |
| *En az tercih edilen bilim dalı tarımsal bilimlerdir.* | Agriculture is the least preferred discipline. | Superlative, relative clause |
| *İlk grafikte erkekler ve kadınları karşılaştırırsak,…* | If we compare men and women on the first graph,… | Present simple, if clause |
| *…bütün kategorilerde erkeklerin daha fazla olduğunu görüyoruz* | …we see that there are more men at all categories. | Comparative, noun clause |

#### 4.2.1.2 The essay task

The essay task requires students to write an argumentative essay in response to a controversial proposal on a social issue or a controversial question. The proposal is presented in the prompt in no more than 30 words. Other than the prompt that specifies the topic, the task does not provide students with any source materials to base their responses; therefore, it is considered to be an independent task. Students are expected to rely on their personal experiences and background knowledge in order to complete the task. As in the graph interpretation task, the topic of the essay task was chosen with an attention of not to privilege or disadvantage any group of students. A controversial proposal on the effects of social media was considered to be relevant to students' interest as different forms of social media such as Twitter, Instagram, Facebook have been very popular for the last few years. Even if not everyone actively uses different forms of the social media, they at least have an idea of their functions and roles in people's lives.

The sub-skills that are required for a successful completion of the task are: 1) presenting a position in response to the question; 2) providing arguments relevant to the position; 3) supporting the arguments with relevant examples; 4) drawing conclusions coherent with the position. These requirements were made clear to the students in the instructions by asking them to write their opinion about the argument with reasons, and support them with examples.

When it comes to the response characteristics, students were expected to write at least 250 words to complete the essay task in 30 minutes. As writing an argumentative task requires relatively extended content generation on the part of the writer in order to expand on the ideas in response to the topic, the essay task has the potential to elicit wider range of lexical items and grammatical structures. Examples of some lexical items and structures from students' responses are presented in Table 7.

Table 7 indicates that students made use of wider range of lexical items and those lexical items differed across various responses, since they all developed different opinions about the subject. Similarly, grammatical structures that students made use of were also more varied. Present simple tense, present perfect tense, modals, noun clauses, relative clauses used to express several language functions were among the structures found in various student responses.

Table 7. Samples from Student Responses on the Essay Task

| Extracts from students' responses | English equivalents | Grammatical structure /Language functions |
|---|---|---|
| *Mesela ben üniversite öğretmenlerini Twitter da takip ediyorum* | For example, I follow my instructors on Twitter. | Present Simple / expressing preference |
| *...sosyal medya günlük yaşamımın bir parçası haline geldi.* | …social media has become a part of my everyday life. | Present perfect tense / reporting a recent phenomenon |
| *…her zaman arkadaşının ne yaptığını, nerde olduğunu öğrenebilirsin.* | …you can always learn what your friends do or where they are. | Modals, noun clause / expressing possibilities |
| *...sosyal medya siteleri insanlara zarar verebilir.* | …social media sites can be harmful for people. | Modals / warning |
| *…farklı kültürleri tecrübe etme imkânı buluruz.* | …we have the chance to experience different cultures. | Noun clause / expressing possibilities |
| *...depresyona giren insanlar artık çevremizde çok.* | …there are now many people around us who get into depression. | There is-there are, relative clause / defining people and objects |
| *...gençler dünyadaki trendleri takip edebiliyor.* | …youngsters are able to follow the trends in the world. | Modals /expressing opportunities |

To summarize, the two task types differed in terms of input stimuli (a graph vs. a stand- alone prompt), discourse mode (descriptive report vs. argumentative essay), the sub-skills necessary for a successful task completion, the cognitive and linguistic demands, the expected length of performance and the time allotted to complete each task. Table 8 summarizes the task characteristics for both tasks.

Table 8. Summary of Task Characteristics

| | The graph interpretation task | The essay task |
|---|---|---|
| Task type | Integrated | Independent |
| Discourse mode | Descriptive report | Argumentative essay |
| Input stimuli | Two bar graphs | A stand-alone prompt |
| Intended CEFR level | B2 | C1 |
| Expected text length | At least 150 words | At least 250 words |
| Time allotted | 20 minutes | 30 minutes |
| Cognitive demands | Graph comprehension, graph interpretation, graph translation | Goal setting, planning, organizing, creation of new knowledge |
| Required Sub-skills | Identifying main trends in the data, selecting important information, making comparisons, organizing information with an order of significance | Presenting a position in response to a question, developing arguments relevant to the position, supporting the arguments with examples, drawing appropriate conclusions |
| Linguistic demands | Task- related vocabulary<br>Relatively limited language structures and functions(i.e. data presentation, comparison data, and source identification) | Wide range of vocabulary<br>Wider range of language structures and language functions (i.e. expressing opinion, reporting recent phenomenon, expressing possibilities /opportunities and warning) |

### 4.2.2 Differences in students' performance in two tasks

Although 47 students participated in the study, 39 responses for each task were used for statistical analysis. Six students did not respond to either the graph interpretation or essay task, and thus completed only one task. The responses of two students could not be assessed as they produced too little texts (fewer than 40 words). These two cases caused missing ratings and were not included in the analysis. For the analysis, corrected data were preferred over raw data. In other words, after all the raters completed their ratings, the data were monitored for significant discrepancies between the two ratings. Five pairs of ratings were found to be discrepant for each task, and they were sent to the raters for a second rating without letting the raters know about their initial ratings. While the initial scores were retained in one of the ten discrepancy cases, the ratings of nine responses were corrected, which accounted for about 6% of the total ratings. Descriptive statistics and the t-test analysis given below show that students scored significantly higher on the essay task.

The descriptive statistics for the scores awarded for each of the four criteria in the scale (i.e., content, organization, language use, vocabulary) for the graph interpretation and essay tasks are presented in Table 9.

Table 9. Descriptive Statistics of Scores for Each Criteria

| | Graph Task | | | | | Essay task | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C | O | LU | V | Total | C | O | LU | V | Total |
| Mean | 4.26 | 4.49 | 4.62 | 4.72 | 17.7 | 4.97 | 5.08 | 5.10 | 5.36 | 20.4 |
| SD | 1.93 | 1.70 | 1.89 | 1.85 | 7 | 1.91 | 1.87 | 2.06 | 2.03 | 7.2 |
| Skewness | 0.40 | 0.16 | 0.57 | 0.57 | 0.5 | 0.21 | 0.21 | 0.14 | 0.04 | 0.1 |
| Kurtosis | -0.89 | -0.45 | -0.79 | -0.62 | -0.9 | -0.85 | -0.88 | -1.20 | -1.21 | -1.3 |

*Note:* C = content, O = organization, LU = language use, V = vocabulary, SD = standard

The means for the graph interpretation task ranging from 4.26 to 4.72 were found to be lower than the means for the essay task varying between 4.97 and 5.36. The means assigned to four criteria for each task suggest that students received higher scores on all the criteria for the essay task than the graph interpretation task. In addition, the means for the criteria were ordered as content, organization, language use and vocabulary from the lowest to the highest for both tasks. The values of skewness and kurtosis were within the acceptable levels (i.e., -2 / +2), suggesting that the scores assigned to four criteria for both tasks seemed to be normally distributed.

A paired samples t-test was conducted to evaluate the difference between the two means of the graph interpretation task (M = 17.7, SD = 7) and the essay task (M = 20.4, SD = 7.2): a statistically significant difference was found between the two means, t (38) = -4.697, p < .000 (two-tailed). The effect size was found to be medium (Cohen's *d* = 0.764).

### 4.2.3 Difference in task difficulty

To investigate the difference in the difficulty levels of the tasks, the variable map and selected statistics from the task measurement report were used. The variable map in Figure 6 displays the measures of student proficiency, rater severity, task difficulty and criterion difficulty on the same logit scale.

The first column in the map shows the logit scale, which is a true interval scale, as opposed to raw scores in which distances between intervals may be different (Park, 2004). The two tasks are compared in terms of their difficulty estimates in the fourth column. The more difficult task appears higher in the column whereas the easier task appears lower. Accordingly, the graph interpretation task appearing higher in the column has a difficulty measure of 0.54 logit and the essay task has a difficulty measure of -0.54 logit. The result indicates that the two tasks are not of equal difficulty, the graph interpretation task being relatively more difficult than the essay task.

```
+------------------------------------------------------------------------+
|Measr|+students |-rater|-task  |-criteria                   | S.1 | S.2 |
|-----+----------+------+-------+----------------------------+-----+-----|
|   7 +          +      +       +                            + (8) + (8) |
|     |          |      |       |                            |     |     |
|     | 11       |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|   6 +          +      +       +                            +     +     |
|     |          |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|   5 +          +      +       +                            +     +     |
|     | 29       |      |       |                            |     |     |
|     |          |      |       |                            |     | --- |
|     |          |      |       |                            | --- |     |
|   4 +          +      +       +                            +     +     |
|     | 19 37    |      |       |                            |     |     |
|     | 2  9  26 |      |       |                            |  7  |  7  |
|     | 32       |      |       |                            |     |     |
|   3 + 16 21    +      +       +                            +     + --- |
|     | 10 30    |      |       |                            | --- |     |
|     | 38       |      |       |                            |     |     |
|     |          |      |       |                            |  6  |  6  |
|   2 +          +      +       +                            +     +     |
|     |          |      |       |                            | --- | --- |
|     |          |      |       |                            |     |     |
|     | 20 31    |      |       |                            |  5  |  5  |
|   1 + 39       +      +       +                            +     +     |
|     | 3  22    |      |       |                            |     |     |
|     | 4  23    |      | graph |                            | --- | --- |
|     | 6  15    | Y    |       | content                    |     |     |
*   0 * 7  12 18 * F    *       * language      organisation *  4  *  4  *
|     | 17       | T    |       | vocabulary                 |     |     |
|     | 28       |      | essay |                            |     |     |
|     | 14       |      |       |                            |     |     |
|  -1 + 13 27    +      +       +                            + --- + --- |
|     | 1  24 34 |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|     | 33       |      |       |                            |     |  3  |
|  -2 +          +      +       +                            +  3  +     |
|     |          |      |       |                            |     |     |
|     | 5        |      |       |                            |     |     |
|     | 35       |      |       |                            |     | --- |
|  -3 +          +      +       +                            + --- +     |
|     |          |      |       |                            |     |     |
|     | 8  36    |      |       |                            |     |     |
|     |          |      |       |                            |     |     |
|  -4 + 25       +      +       +                            + (1) + (1) |
|-----+----------+------+-------+----------------------------+-----+-----|
|Measr|+students |-rater|-task  |-criteria                   | S.1 | S.2 |
+------------------------------------------------------------------------+
```

Figure 6. FACETS summary (student proficiency, rater severity, task and criteria difficulty) Note: S.1 = Scoring rubric used for graph task; S.2 = Scoring rubric used for essay task

A summary of selected statistics included in the task measurement report is provided in Table 12. The full report is presented in Appendix G.

Table 12. Summary of Statistics Included in the Task Measurement Report

| Task | Difficulty Measure | Standard error | Infit Mean-Square Index | Outfit Mean-Square Index |
|---|---|---|---|---|
| Graph | 0.54 | 0.07 | 1.08 | 1.09 |
| Essay | -0.54 | 0.07 | 0.83 | 0.80 |
| Mean | 0.00 | 0.07 | 0.95 | 0.94 |
| S.D. | 0.51 | 0.00 | 0.12 | 0.14 |

*Note:* Reliability of separation index = 0.98; separation index = 10.46; fixed chi-square: 117.3, *df* = 1, *p* = .00

The first column shows the two tasks identified by task types. The difficulty measures shown in the second column indicate that it was harder for students to receive high ratings on the graph interpretation task than on the essay task (i.e., the graph interpretation task has a difficulty measure of 0.54 logit, while the essay task has a difficulty measure of -0.54 logit).

The task separation index (10.46) and the reliability of separation index (0.98) indicate that the two tasks were not of equal difficulty. This finding was further confirmed by the significant the fixed chi-square test value.

The infit and outfit mean-square values for the two tasks were within the acceptable range, indicating that the ratings for the two tasks showed sufficient fit to the measurement model.

### 4.2.4 Separation of students into distinct categories

To examine if the tasks reliably separated students into distinct categories, the variable map, and selected statistics from the student measurement report were used.

The second column in the variable map (See Figure 6) shows estimates of student proficiency.  Each number represents one student, and higher scoring students appear at the top of the column whereas lower scoring students appear at the bottom, logit 0 being the average. The distribution of student proficiency measures is quite wide, ranging from a high of 6.48 logits to a low of -3.92 logits. If the two outlier students (Student #29, Student #11) are excluded, the student proficiency measures appear to be normally distributed, with measures ranging between 3.87 logits and -3.92 logits. However, one can also argue that there are more high achieving students (above 0 logit) than low achievers, and thus the distribution of student proficiency measures is somewhat negatively skewed.

A summary of selected statistics included in the student measurement report is provided in Table 10. The student proficiency measures ranged from -3.92 to 6.48 logits. The mean of the proficiency measures was 0.70 logit (SD = 2.44). The student separation index was 9.46, and the reliability of student separation was 0.98.  The separation index is an estimate of the number of distinguishable levels of proficiency among the students. The separation index of 9.46 indicates that there were about nine statistically distinct strata among the 39 student proficiency measures.

Table 10.  Summary of Results for Students (*N* = 39)

| | |
|---|---|
| Mean of the proficiency measures | 0.70 |
| Standard deviation of the proficiency measures | 2.44 |
| Student separation index | 9.46 |
| Reliability of student separation | 0.98 |
| Fixed (all same) chi-square | 1771.8 (*df* = 38, *p* = .00) |

The reliability of the student separation is the Rasch equivalent of KR20 or Cronbach Alpha statistics (O'Sullivan, 2005). A reliability coefficient of 0.98 indicates that the raters' ratings on the two tasks reliably separated students into different levels of proficiency. It also suggests that those ratings did not show evidence of central tendency error. The significant chi-square test value indicates that the null hypothesis (all students were equally proficient) should be rejected.

Fit statistics were examined in order to identify students who displayed unusual rating profiles. There are different suggestions for setting the lower and upper limits of infit and outfit mean-square statistics. In this study, the range between 0.5 and 1.5 was accepted as the standard as Linacre (as cited in Eckes, 2009) suggested. Based on these criteria, there were three misfitting students. However, if we take into consideration the fact that values between 1.5 and 2 are considered less productive for measurement, but not distorting, particularly in the context of low-stakes assessments by Linacre (as cited in Sudweeks et al., 2005), the infit mean-square values for  student #22 were within acceptable levels. Table 11 presents the fit statistics and the rating patterns for these students. Table 11 shows that for the graph interpretation task, Student #16 received unexpectedly low ratings from both Rater #1 and Rater #3 for content and an unexpectedly low rating from Rater #3 for organization, while those two raters gave him much higher ratings for language use and vocabulary. His high ratings on all four criteria for the essay task seem to be consistent with those other high ratings that he received for the graph interpretation task. However, his unexpectedly low ratings for content and organization indicate that although the two raters judged the student as highly proficient when completing the essay task, the student was probably not as familiar with the content or organizational requirements of the graph interpretation task.  An analysis of this student’s response on task 1 confirmed the fact that task unfamiliarity was the reason for the low ratings the student received on the graph interpretation task.

Table 11. Rating Patterns and Fit Indices for Misfitting Students

Ratings received by Student #16
(Infit Mean-Square Index = 3.11 Proficiency Measure = 3.03)

| Task1(graph) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #1 | 3 | 5 | 8 | 8 |
| Rater #3 | 2 | 4 | 7 | 7 |

| Task2(essay) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #1 | 7 | 7 | 8 | 8 |
| Rater #2 | 8 | 8 | 8 | 8 |

Ratings received by Student #22
(Infit Mean-Square Index = 1.59 Proficiency Measure = 0.63)

| Task1(graph) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #1 | 1 | 4 | 4 | 4 |
| Rater #2 | 2 | 4 | 5 | 4 |

| Task2(essay) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #3 | 6 | 5 | 6 | 6 |
| Rater #2 | 6 | 5 | 6 | 5 |

Ratings received by Student #35
(Infit Mean-Square Index = 1.96 Proficiency Measure = -2.82)

| Task1(graph) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #1 | 1 | 4 | 3 | 4 |
| Rater #2 | 2 | 3 | 4 | 3 |

| Task2(essay) | Content | Organization | Language use | Vocabulary |
|---|---|---|---|---|
| Rater #3 | 1 | 2 | 2 | 2 |
| Rater #2 | 2 | 2 | 3 | 3 |

Similarly, for the graph interpretation task, Student #22 received unexpectedly low ratings from both Rater #1 and Rater #2 for content, whereas he received ratings of 4, 5, or 6 for the other three criteria. An analysis of the student's response indicated that the student confused the phrase *doktora yapan bireyler* (people holding doctoral degrees) with *doktora giden bireyler* (people who go to the doctor) as these two phrases are very similar in Turkish. On the essay task, he also received ratings of 5 and 6 for all four criteria. Student #35, on the other hand, seems to misfit because for the graph interpretation task, he received higher ratings than expected for organization, language use and vocabulary, given the other ratings that he received.

The pinpointing of unexpectedly low or high ratings is important in that those ratings can provide useful information to determine whether each student's proficiency measurement is an indication of a valid and trustworthy measurement of that student's abilities. Therefore, it is important to review the rating profiles of misfitting students and identify whether unexpectedly high or low ratings are related with construct relevant or construct irrelevant variables before issuing their score reports.

### 4.3 Research question 2: How reliably does the rating scale function for this specific group of raters?

In order to investigate the second research question, selected statistics from many-facet Rasch measurement analysis were reported. Specifically, selected statistics from the criteria measurement report, rater point biserial correlation indices and rater fit indices from the rater measurement report and selected statistics from the category statistics were used to examine the effectiveness of the rating scales used in the present study.

### 4.3.1 Criteria

A summary of selected statistics included in the criteria measurement report is provided in Table 13 (See Appendix G for the full report).

Table 13. Summary of Statistics Included in the Criteria Measurement Report

| Criterion | Difficulty Measure | Standard Error | Infit Mean-Square Index | Outfit Mean-Square Index |
|---|---|---|---|---|
| Content | 0.31 | 0.10 | 1.40 | 1.34 |
| Organization | 0.03 | 0.10 | 0.93 | 0.99 |
| Language Use | -0.07 | 0.10 | 0.82 | 0.81 |
| Vocabulary | -0.28 | 0.10 | 0.67 | 0.64 |
| Mean | 0.00 | 0.10 | 0.95 | 0.94 |
| S.D. | 0.21 | 0.00 | 0.27 | 0.26 |

Note: Reliability of separation index = 0.78; separation index = 2.82; fixed chi-square:17.9, *df:*3, *p* = .00

The first column presents the four criteria that the raters used to evaluate the students' responses. The second column displays the difficulty measures for the criteria. The hardest criterion to get high ratings on was content (0.31 logit). By contrast, the easiest criterion to get high ratings on was vocabulary (-0.28 logit). The difficulty measures for organization (0.03 logit) and language use (-0.07 logit) were very similar.

The criteria separation index (2.82) and the reliability of criteria separation (0.78) suggest that the four criteria differed somewhat in difficulty. Among the four criteria, there were nearly three statistically distinct levels of difficulty, which may provide evidence for the absence of halo effect.

The infit and outfit mean-square values for the criteria were within the acceptable range of 0.5 to 1.5, indicating that there were no overfitting or misfitting criteria. The fact that there were no overfitting criteria suggests that the four criteria were not scored too similarly, and the fact that there is no misfitting criteria provides evidence for psychometric unidimensionality of the four criteria, suggesting that they might be all associated with the same dimension (Eckes, 2009). In other words, ratings on one criterion agree well with the ratings on other criteria, leading to a single pattern of proficiency across all four criteria (Park, 2004).

### 4.3.2 Inter-rater & intra-rater reliability estimates

To measure inter-rater reliability, FACETS provides two measures of rater reliability: the rater point biserial correlation index and the percentage of exact rater agreement. The rater point biserial correlation index is a measure of how similar are the raters in their rankings of students, and the percentage of exact agreement shows the percentage of how many times the raters assigned exactly the same score as another rater (Knoch, 2007). Table 14 provides the summary of these two rater reliability measures.

Table 14. Summary of Rater Reliability Measures

| Rater | Rater Point Biserial Measure | Percentage of Exact Agreement |
|---|---|---|
| F | 0.90 | 48.1 % |
| T | 0.91 | 42.7 % |
| Y | 0.88 | 37.5 % |

The first column in Table 14 shows raters IDs. The second column displays point biserial correlation indices for raters. Myford and Wolfe (2003, p. 416) use the term “single rater-rest of raters correlations” for this type of correlation index, which means that each correlation index indicates the correlation measure of one rater with the other two raters within this group of raters. Accordingly, the single rater-rest of rater correlations seem to be substantial, which were 0.90, 0.91, and 0.88 for Rater

#F, Rater #T and Rater #Y, respectively. This suggests that there was a significant level of reliability between the raters. The third column indicates that Rater #F has the highest exact agreement percentage (48.1%), suggesting that Rater #F awarded exactly the same scores 48.1% of times as the other raters under the same conditions, while Rater #Y has the lowest agreement percentage (37.5%).

For intra-rater reliability, rater infit and outfit mean square values are provided by the rater measurement report, which is presented in section 4.4.1.

### 4.3.3 Rating scale (category statistics)

In order to examine whether eight-point rating scales which were used to score students' responses for graph interpretation and essay tasks functioned as intended, a summary of selected category statistics are presented in Table 15 (See Appendix G for the full report).

Table 15. Category Statistics for Rating Scales

| | Graph | | | | Essay | | | |
|---|---|---|---|---|---|---|---|---|
| Category | Average Measure | Outfit | Threshold | SE | Average Measure | Outfit | Threshold | SE |
| 1 | -3.53 | 1.4 | | | -2.77 | 1.0 | | |
| 2 | -2.98 | 1.2 | -5.6 | 0.43 | -2.72 | 0.6 | -5.85 | 1.01 |
| 3 | -1.36 | 1.1 | -3.11 | 0.24 | -1.17 | 0.9 | -2.61 | 0.28 |
| 4 | -0.46 | 1.2 | -0.74 | 0.18 | 0.09 | 0.8 | -0.82 | 0.20 |
| 5 | 0.68 | 0.9 | 0.94 | 0.22 | 0.95 | 0.9 | 0.72 | 0.18 |
| 6 | 2.19 | 1.2 | 1.46 | 0.24 | 2.09 | 0.8 | 1.82 | 0.22 |
| 7 | 3.16 | 0.8 | 3.01 | 0.22 | 3.75 | 0.6 | 2.90 | 0.23 |
| 8 | 4.14 | 0.9 | 3.50 | 0.28 | 4.62 | 1.0 | 3.83 | 0.22 |

Note: Thresholds = Rasch-Andrich thresholds, SE = Standard error

The first column in Table 15 shows category labels as appeared in scoring scales ranging from 1 and 8. The second and the sixth column indicate the average student proficiency measure by rating scale category. Linacre (2002) suggests that the average measures should advance monotonically as the categories increase. It seems in the table that the average measures for both scales increase as the categories increase. For example, the first rating scale used for the graph interpretation task increased from -3.53 to 4.14 as the categories increased from 1 to 8, suggesting that the categories were ordered appropriately and meaningfully.

Outfit mean-square indices are presented in the third column for the scoring scale used for the graph interpretation task and in the seventh column for the scale used for the essay task. Outfit mean-square index is another indicator of rating scale functionality as suggested by Linacre (2002). FACETS computes average student proficiency measure and an expected student proficiency measure. The larger the discrepancy between the average and expected measures, the larger the outfit mean-square index will be (Eckes, 2009). Linacre (2002) suggests that outfit mean-square index should be less than 2.0 as a high value of mean-square related to a category provides evidence for the fact that the category has been used in unexpected contexts. As shown in Table 15, all the outfit mean-square values were less than 2.0, which suggests that the categories for both rating scales seemed to function as intended.

The category thresholds can also provide information on the quality of a rating scale. The columns 4 and 8 show the category thresholds for each rating scale. Just as the average measures, it is expected that these thresholds advance monotonically with categories. When they do not, they are disordered, suggesting low probability of occurrence of certain categories due to the rating behavior in which those categories are employed (Linacre, 2002). Table 15 shows that threshold measures advance monotonically as the categories increase (i.e., from -5.6 to 3.50 for the first scale, and from -5.85 to 3.83 for the second scale).

### 4.4 Research question 3: To what extent is the quality of ratings influenced by rater effects?

This question was examined in terms of rater severity, rater inconsistency, central tendency and halo effect through selected statistics from the rater measurement report provided by many-facet Rasch measurement analysis.

#### 4.4.1 Rater severity

The rater measurement report provides useful information regarding rater behaviors. It reports a measure of the level of severity each rater exercised, as well as measures of each rater's ability to use the rating scales in a consistent manner when evaluating multiple students' responses. Table 16 provides a summary of some of the statistics included in that report (i.e., rater severity measures, standard errors of the severity measures, infit and outfit mean-square indices). The full report is provided in Appendix G.

Table 16. Summary of Statistics Included in the Rater Measurement Report

| Rater ID | Severity Measure | Standard error | Infit Mean-Square Index | Outfit Mean-Square Index |
|---|---|---|---|---|
| F | -0.03 | 0.08 | 0.92 | 0.90 |
| T | -0.15 | 0.09 | 0.82 | 0.86 |
| Y | 0.18 | 0.09 | 1.14 | 1.09 |
| Mean | 0.00 | 0.09 | 0.96 | 0.95 |
| S.D. | 0.14 | 0.00 | 0.13 | 0.10 |

Note: Reliability of separation index = 0.61, separation index = 2.01

The first column shows the rater IDs. The second column shows that the difference between the severity measures of the most severe (Rater #Y) rater and the most lenient (Rater #T) rater was 0.33 logits, indicating that the three raters appeared to exercise similar levels of severity when rating students' responses. The rater separation index indicates the number of statistically distinct groups in terms of rater severity. Thus, the separation index of 2.01 suggests that there were about two statistically distinct strata of rater severity within this small group of raters. The reliability of rater separation index indicates how different the raters are in their severity measures unlike inter-rater reliability, which is a measure of how similar the raters are in their severity measures (Eckes, 2009, p. 20).  In other words, when raters display similar measures of severity, the reliability of separation index is expected to be close to 0. Therefore, a low separation reliability index is desirable for raters. The rater separation reliability index for this analysis was 0.61, indicating that the raters differed somewhat in their severity.

### 4.4.2 Rater inconsistency

To examine rater inconsistency, rater fit indices were used. The fourth and fifth columns in Table 16 show rater fit statistics. One examines rater fit statistics to determine whether raters used the rating scales in a consistent manner (Eckes, 2009). The infit and outfit mean-square values for all three raters were within the range of 0.5 and 1.5, which means that none of them were misfitting. That is to say, all of the raters were self-consistent in their ratings.

### 4.4.3 Central tendency

Central tendency was examined through rater fit indices and student separation index. The fact that there were no overfitting raters (i.e., no infit mean-square values lower than 0.5) suggests that the raters did not tend to overuse certain (generally middle) scale categories, which could lead the raters to appear as too consistent. An overfitting rater is one who has assigned ratings that are closer to the expected ratings than the measurement model predicted, which was not the case with this particular group of raters (Knoch, 2007). Student separation index was presented in section 4.2.4.

### 4.4.4 Halo effect

See section 4.3.1 for the results of the criteria separation index statistics, which was used to examine possible involvement of halo effect in the ratings.

## 4.5 Research question 4: How dependable are the scores assigned to the examinees?

The results of generalizability theory analysis were reported to examine the dependability (reliability) of the test scores. G theory analysis consisted of two stages: G-studies and D studies.

### 4.5.1 G-studies with p x t x r’ design

G studies were conducted using fully crossed p x t x r’ design in order to estimate main effect sizes of persons, tasks, ratings and their interactions on the overall score variance. The analysis was conducted for the four scoring criteria separately in order to examine the contribution of each criterion to the overall score variability. The estimated variance components and the percentage of total variance explained by each source of variance for the four criteria are presented in Table 17.

Table 17. Variance Components for p x t x r' Design

| Sources of Variation | Variance Component | | | | Percent of Total Variation (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | C | O | LU | V | C | O | LU | V |
| P | 2.48 | 2.26 | 3.5 | 3.16 | 62.3 | 68.1 | 80 | 79 |
| T | 0.31 | 0.14 | 0.16 | 0.22 | 7.8 | 4.4 | 4.0 | 5.6 |
| R' | -0.02 | 0.03 | 0.00 | 0.00 | 0.0 | 1.0 | 0.0 | 0.1 |
| PT | 0.72 | 0.47 | 0.34 | 0.13 | 18.0 | 14.2 | 8.8 | 3.4 |
| PR' | 0.02 | 0.03 | 0.00 | -0.13 | 0.5 | 0.9 | 0.1 | 0.0 |
| R'T | 0.03 | 0.01 | 0.00 | 0.09 | 0.7 | 0.3 | 0.0 | 0.3 |
| PTR' | 0.43 | 0.037 | 0.28 | 0.45 | 10.8 | 11.1 | 7.1 | 11.3 |

Note: P = person, T = task, R'= rating, PT = person by task, PR'= person by rater, R'T = rating by task, PTR' = person by task by rating; C = content, O = organization, LU = language use, V = vocabulary

Based on the design of this study, seven different components of variance were obtained: persons (P), ratings (R'), tasks (T), person-by-task (PT), person-by-rating (PR'), and person-by-task-by-rating (PTR'), representing a triple interaction between persons, tasks, ratings, along with all other unexplained sources of variations. Persons (P) stand for students in the context of this study and they constitute the object of measurement, not a source of error (Webb & Shavelson, 2005). Therefore, variance associated with persons represents systematic individual differences in terms of writing proficiency, and it is ideally expected to be larger than any other sources of variance.

Table 17 indicates that the largest percentage of total variance was explained by real differences among persons (students) for all the four criteria, the percentages of variance ranging from 62.3% to 80%. The scores associated with language use and vocabulary criteria seem to be the most dependable, as 80% and 79% of the total variance was accounted for by the variance related to students for these criteria.

The estimated variance of task (T) reflects differences in the difficulty levels of tasks. Table 17 shows substantial task main effects across all the criteria, content having the largest proportion of variance (7.8%). This suggests that students differed in the ways they responded to the two tasks, and this difference was the most obvious for content criterion. Similarly, considerable effect of variance was explained by person-by-task interaction (i.e., content = 18%, organization =14.02%, language use = 8.8%, vocabulary = 3.4%). The relatively large values of person-by-task interaction suggest that the rank-ordering of a significant number of students was not consistent across tasks (Lee, 2006). This difference was largest for content, and smallest for vocabulary.

The main effect for ratings (R’) shows the degree of consistency between the two ratings assigned to the same performance by different raters. Table 17 shows that there was no significant variance associated with rating main effect though ratings differed somewhat for organization (explaining 1.0% of the total variance). Small values of person-by-rating interaction (i.e., content = 0.5%, organization = 0.9%, language use = 0.1%, vocabulary = 0%) provide further evidence for the fact that the rank-ordering of students were somewhat similar across the first and second ratings for all the criteria. The relatively large three-way, person-by-task-by-rating interaction suggests that the observed person-by-task interaction was not consistent across various ratings. (Sudweeks et al., 2005).

The large variance percentages explained by task and person-by-task interaction suggest that the task facet contributed substantially to variability in observed scores whereas the rating facet did not seem to have a significant effect on score variability, which was evidenced by ignorable effect sizes of rating and person-by-rating interaction. Therefore, it is suggested that any generalizations about students' relative ranking based on only one of the tasks could lead to misinterpretations about students' ability and would not be dependable (Sudweeks et al., 2005).

### 4.5.2 D-studies

It is possible to obtain a number of scenarios to examine the optimal number of tasks and ratings for maximizing score dependability by conducting D-studies. D-studies are conducted using the variance components produced in G- studies (Xi, 2007). The G theory analysis provides two types of reliability coefficient: a generalizability coefficient (G coefficient) for relative decisions and an index of dependability (phi coefficient) for absolute decisions (Webb & Shavelson, 2005). Relative decisions are concerned with rank ordering the students (norm-referenced) in such situations as selecting individuals for college or job. Absolute decisions are, on the other hand, related with individual performance "…regardless of how others perform on the test" as in criterion referenced tests (Webb & Shavelson, 2005, p. 604). Changes in the G coefficients were examined for this study rather than the phi coefficients as the purpose of the study was to rank order the students based on their proficiency levels (i.e., relative decision). Table 18 shows the reliability coefficients (G coefficients) estimated in the various D-studies for analytic scores using p x T x R' design. Sixteen different scenarios with tasks and ratings increasing from 1 to 4 were presented in Table 18.

Table 18. Changes in G- Coefficients of the Four Analytic Scores in D Studies

| | | The G coefficients | | | |
|---|---|---|---|---|---|
| No. of tasks | No. of ratings | C | O | LU | V |
| 1 | 1 | 0.68 | 0.72 | 0.83 | 0.84 |
| 1 | 2 | 0.73 | 0.77 | 0.87 | 0.90 |
| 1 | 3 | 0.74 | 0.79 | 0.88 | 0.92 |
| 1 | 4 | 0.75 | 0.80 | 0.88 | 0.93 |
| 2 | 1 | 0.81 | 0.83 | 0.91 | 0.92 |
| [4]2 | 2 | 0.84 | 0.87 | 0.93 | 0.95 |
| 2 | 3 | 0.85 | 0.88 | 0.93 | 0.96 |
| 2 | 4 | 0.86 | 0.89 | 0.94 | 0.96 |
| 3 | 1 | 0.86 | 0.88 | 0.94 | 0.94 |
| 3 | 2 | 0.89 | 0.91 | 0.95 | 0.96 |
| 3 | 3 | 0.89 | 0.92 | 0.96 | 0.97 |
| 3 | 4 | 0.90 | 0.92 | 0.96 | 0.97 |
| 4 | 1 | 0.89 | 0.90 | 0.95 | 0.96 |
| 4 | 2 | 0.91 | 0.93 | 0.96 | 0.97 |
| 4 | 3 | 0.92 | 0.93 | 0.97 | 0.98 |
| 4 | 4 | 0.92 | 0.94 | 0.97 | 0.98 |

Note: C = content, O = organization, LU = language use, V = vocabulary

One observation of these scenarios is that G-coefficients increased when more tasks and ratings were used, but the relative effect of the number of ratings on the score reliability was smaller than the effect of the number of tasks for each criterion. For example, when the number of tasks was increased from 1 to 3 in a single rating scenario, the G- coefficients changed dramatically from 0.68 to 0.86 for content, from 0.72 to 0.88 for organization, from 0.83 to 0.94 for language use and from 0.84 to 0.94 for vocabulary. Increasing the tasks from 3 to 4, however, did not lead to

[4] Two ratings two tasks scenario as used in the main study.

much difference. There was an increase of 0.3, 0.2, 0.1 and 0.2 in the G-coefficients for content, organization, language use and vocabulary, respectively.

Increasing the number of ratings from 1 to 2 contributed to the reliability coefficients of the analytic scores whereas the impact of three or four ratings was small. For example, the G coefficients increased from 0.68 to 0.73 for content, from 0.72 to 0.77 for organization, from 0.83 to 0.87 for language use, and from 0.84 to 0.90 for vocabulary in a one task double rating scenario. However, the G coefficient changed only from 0.73 to 0.74, from 0.77 to 0.79, from 0.87 to 0.88 and from 0.90 to 0.92 for content, organization, language use and vocabulary in a 1 task 3 ratings scenario, respectively. These findings related to various scenarios were not surprising considering the relatively large main effect of tasks and small main effect of ratings.

## 4.6 Conclusion

The chapter provided the results from the qualitative and quantitative analyses that were used to investigate the research questions. Qualitative analysis included detailed description of the task characteristics in terms of the cognitive demands and contextual features of the graph interpretation task and the essay task. Quantitative analysis consisted of descriptive statistics, selected statistics from the many-facet Rasch analysis, and G theory analysis. An extended discussion of these results will be presented in the next chapter.

# CHAPTER 5

# DISCUSSION

The analyses that were conducted with the aim of providing evidence for scoring validity as well as cognitive and contextual validity for the newly developed TFL writing tests were reported in the previous chapter. A comprehensive discussion of the results reported from the analyses will be presented in accordance with research questions in this chapter.

Research question 1: How do the graph interpretation and essay tasks differ from each other in terms of cognitive and linguistic demands?

In order to test students' academic writing skills of Turkish, two different task types were proposed in the study. One of the tasks is a traditional argumentative essay task, in which students are expected to present a position in response to a given proposal or a controversial question and develop arguments to support their position. The essay task was chosen because previous studies indicate that essays are one of the most frequently practiced task types at university (Hale et al., 1996, as cited in Paltridge, 2004; Moore & Morton, 2005). Moore and Morton (2005), for example, investigated the written genre and text type requirements of undergraduate and postgraduate students in two Australian universities and found that the essay was the most common genre, accounting for 60% of the complete set of tasks assigned at both undergraduate and post graduate level. Therefore, the essay task is considered to exemplify the type of writing tasks frequently practiced in academic environment, the domain which is determined as the target language use domain for the study.

However, this type of independent (stand-alone) essay tasks have been criticized by several researchers as no source of information is provided to students except for the prompt that outlines the topic (i.e., Cumming et al., 2000; Gebril, 2009; Plakans, 2007; Weigle, 2002). It is argued that university students generally rely on different source materials such as lecture notes, course books, and class discussions, which provide them with necessary background information in order to carry out academic writing tasks required in their courses, whereas a traditional, stand-alone essay will require students to rely on their background knowledge and personal experiences. Therefore, such a task would fail to reflect what students actually do in real-life academic settings. For this reason, an integrated graph interpretation task in which students were required to describe the information provided in the form of a bar graph was used along with the independent essay task, as graph interpretation tasks are often considered an integrated task (Ahmadi & Mansoordehghan, 2015; Yang, 2012; Yu, Rea-Dickens & Kiely, 2007) as well as more frequently used reading-to-write or listening-to-write integrated tasks. Therefore in this study, greater degrees of the authenticity and generalizability of tasks to the target language use domain tasks were ensured by selecting tasks that exemplify two different task types used in academic settings. On the other hand, using two different task types raised the questions of whether these task types differ in cognitive and linguistic demands they place on the test takers and whether the students' performance on these tasks was comparable. In other words, are these two task types associated with the same writing construct or are they tapping into different constructs of writing ability?

A qualitative analysis of the cognitive and contextual characteristics of tasks along with language samples from the students' responses revealed that two task types varied systematically in terms of discourse mode, input stimuli, intended CEFR level, expected length of student response, the time allotted to complete each task, and cognitive and linguistic demands. Accordingly, describing the information given in a bar graph was considered to be cognitively less demanding than presenting a position and developing relevant arguments in response to a controversial proposal and supporting the arguments with relevant ideas and examples. This assumption was further supported with cognitive models of knowledge telling and knowledge transforming proposed by Breiter and Scardamalia (1987, as cited in Grabe and Kaplan, 1996). Based on these models, it can be argued that the graph interpretation task requires more knowledge telling skills as the task requires the written description of the existing visual data. Writing an argumentative essay, on the other hand, tends to involve such processes as an analysis of the problem, goal setting and creation of new knowledge, which are related with knowledge transforming skills. Therefore, it was expected that students would receive higher scores on the cognitively less demanding task (the graph interpretation task) than on the essay task.

However, contrary to the research expectation, the quantitative analysis used to examine the effect of task types on students' performance revealed that students scored higher on the essay task, which was assumed to be cognitively more demanding. Descriptive statistics indicated that the mean scores assigned to the essay task responses were higher across all of the four scoring criteria (i.e., content, organization, language use and vocabulary), and the t-test showed that the difference was statistically significant with a medium effect size. Despite the fact that the finding was unexpected, it seems to be in line with the findings of Crowhurst (1980, as cited in Hout, 1990) and Quellmalz et al. (1982, as cited in Hout, 1990). In both studies, the researchers investigated the effect of discourse mode on the students' performance. Crowhurts' study revealed that students produced longer T units in argumentative essays than in narrative essays, and thus received higher scores on the former task. Although narrative essays are considered to require knowledge telling skills (Grabe and Kaplan, 1996), and thus cognitively less demanding than argumentative essays, students in Crowhurts' study demonstrated a better performance on the latter task. Similarly, Quellmalz et al. found that students performed better on cognitively more demanding task (expository tasks) than on the narrative tasks.

The finding may also provide evidence for the model of Cognition Hypothesis proposed by Robinsons (2001) in an attempt to explain task difficulty. The model of Cognition Hypothesis basically suggests that increasing the difficulty of tasks may lead to higher levels of linguistic complexity, lexical variation and accuracy, as opposed to the Limited Attentional Capacity Model proposed by Skehan and Foster (2001, as cited in Harsch & Rupp, 2011), which claims that increasing the level of task difficulty will lead to linguistically less complex and more erroneous output. Based on these findings, it might be suggested that more complex tasks have the potential to elicit better performance. However, it should be noted that this conclusion is drawn based solely on the test scores, rather than analysis of students' responses in terms of linguistic complexity, lexical variation and accuracy. In order to be more accurate in making such a claim, a comprehensive analysis of students' responses is necessary.

This finding was further supported by the analysis of many-facet Rasch analysis (FACETS). When the variable map generated by many-facet Rasch Analysis (FACETS) was examined, the graph interpretation task was found to be more difficult for students to receive high scores in than the essay task. The selected statistics of task measurement report (i.e., high values of the task separation index, the reliability of separation index, and the fixed chi-square test) further confirmed that the two tasks are not of equal difficulty, the graph task being more difficult for the group of students participated in the study. A difference in task difficulty was expected by the researcher, but the difference turned out to be in the opposite direction. In other words, the graph interpretation task, which was assumed to be at B2 levels was more difficult for students than the essay task that was assumed to be at C1 level to get high scores. A similar result was found in Ellis, Chong and Choy's (2013) study, in which they examined writing proficiency of student teachers at the National Institute of Education in Singapore through IELTS exam. Ellis et al. found that student teachers received lower mean scores on Task 1 (the graph interpretation task) than on Task 2 (the essay task). One possible explanation of this finding is that such types of graph interpretation tasks may require special training and practice as students may not be familiar with the requirements of writing a report based on visual data as much as writing a traditional essay. Ellis et al. (2013) attributed student teachers' higher performance on Task 2 to more school practice in this genre (argumentative essay) Ahmadi and Mansoordehghan (2015) investigated the effect of task types and prompts on test performance by comparing test takers' performance on IELTS Academic Writing Task 1 and Task 2. They stated that it was difficult for test takers to complete Task 1 before they received instruction. The test takers were confused about, for example, how to start their descriptions or what points on the graph to focus on. However, after receiving instructions and training for Task 1, the test takers found Task 1 easier than Task 2. Test takers further stated that knowing certain strategies, lexical items and grammatical structures could help them to be successful on different versions of the same task type. The findings of the present study along with the findings of Ahmadi and Mansoordehghan (2015) and Ellis et al. (2013) suggest that graph interpretation tasks may mask students' actual proficiency due to the task unfamiliarity or a lack of task related vocabulary, and thus special training for this type of tasks is necessary in order to eliminate or at least minimize construct irrelevant variance that might involve in such type of tasks.

Selected statistics from the student measurement report indicated that the two tasks were successful in reliably separating students in different levels of proficiency. The student separation index showed that there were about nine distinct strata of student proficiency, suggesting that both tasks functioned as intended. In the group, only three students displayed unusual rating profiles (misfitting profiles). Analysis of these misfitting students revealed that two of these students' unexpected ratings were associated with the graph interpretation task. The responses of these two students were examined in order to pinpoint what went wrong with these two responses that led to unexpected ratings. Accordingly, student #16 seems to have responded Task 1 as if it was an opinion essay. In other words, instead of simply describing the data provided in the graphs, he attempted to explain the reasons why more men appeared to hold doctoral degrees than women, and why the majority of people preferred medical sciences for their doctoral studies. Although he had a sophisticated use of grammatical structures and lexical items, he failed to meet the requirements of content and organization related with the graph interpretation task. Therefore, he received unexpectedly low ratings on content and organization from both raters. Student #22, on the other hand, seems to have received unexpectedly low ratings on content for the graph interpretation task not because he failed to describe the information in the graphs, but because he misunderstood the topic and wrote his report to his understanding of the task. In other words, he did not have much problem with identifying and describing main trends or comparing the information provided in two different graphs; however, he confused the phrase *doktora yapan bireyler* (people holding doctoral degrees) with *doktora giden bireyler* (people who go to the doctor). Therefore, he was scored as off-topic content wise. The response analysis of student #16 provided further evidence for the existence of unfamiliarity of content and organizational requirements of the graph interpretation task. As Ahmadi and Mansoordehghan (2015) suggested in their study, special instruction and training may lead to better performance on this type of task since certain strategies and a somewhat limited range of lexical items and grammatical structures seem to apply for different versions of the same task type. On the other hand, the fact that majority of the students received lower scores on all the criteria compared to their scores for the essay task may indicate that these two task types are associated with different sub-constructs of academic writing ability. Therefore, it may be argued that an attainment of one does not necessitate the existence of the other.

Research question 2: How reliably does the rating scale function for this specific group of raters?

Rating scales that are used to score students' responses play a critical role in obtaining reliable measures of students' abilities. In other words, a rating scale is one of the factors that can affect validity of our decisions based on the writing test scores assigned by the raters. Therefore, an investigation of how reliably a rating scale functions for its intended purpose is necessary if one desires to achieve higher degrees of validity.

Quantitative analysis was used to examine the reliability of the rating scales. East (2009) citing Cherry and Meyer (1993) notes that "the more pieces of information available, the more reliable will be the conclusions drawn from the data" (p. 92). In other words, multiple ways used to gather evidence for the reliability of the rating are likely to increase the trustworthiness of the conclusions. One of these ways is to examine the results of the criteria measurement report generated by many-facet Rasch analysis. The statistics of criteria measures showed that the most difficult criterion was content, whereas vocabulary was the easiest criterion for the students to get high ratings. In other words, the students had the tendency to score higher on vocabulary than on the other criteria, and they scored lowest on content. Organization and language use were of similar difficulty although they received slightly higher scores on language use. This finding suggests that learners of TFL may have differing proficiency levels in different aspects of writing ability, and the analytic rating scales used in the study were able to reflect the uneven profile of L2 learner's writing proficiency just as Weigle (2002) points out with regard to advantages of using an analytic rating scale.  In addition, the finding of the study seems to confirm previous research (i.e. Bacha, 2001).  Bacha (2001) found out that the analytic ratings assigned for different components of writing were significantly different from each other in her study with L1 Arabic students of English. As noted before, the finding of the present study might be useful in diagnosing specific areas in which TFL students have problems in academic writing tasks. It seems that TFL students had the necessary vocabulary repertoire and linguistic structures to complete the academic tasks, but they generally had problems generating ideas on the given subject (content) and organizing their ideas as required by each task (organization). This problem was more salient with the graph interpretation task. The analytic scores helped to understand that some students were unfamiliar with the content requirements of the graph writing task despite the fact that they had a good control of grammatical structures and vocabulary.

The criteria separation index generated by the criteria measurement report indicated nearly three statistically distinct levels of difficulty among the four criteria.

This finding suggests that raters were able to distinguish at least three criteria in the rating scales, and thus the rating scales functioned analytically as intended. This finding may provide support for the effectiveness of empirically developed rating scales as suggested by various researchers (i.e., Knoch, 2007; Turner & Upshur, 2002). The descriptors of empirically-based rating scales are based on the analysis of actual student responses; therefore, they are argued to be more discriminating and explicit in terms of their level descriptors than intuitively developed rating scales (Turner & Upshur, 2002). Knoch (2007) compared an intuitively developed analytical scale (the existing scale) with an empirically developed rating scale (the pilot scale). The findings of her study revealed that the existing intuitively developed rating scale functioned well but in a holistic manner. In other words, despite the fact that raters used an analytical scale, they resorted to their impressionistic judgments in their ratings as indicated by the scores being generally the same across various traits. On the other hand, that was not the case with the pilot scale, and the raters seemed to have used the scale analytically. Similarly, the fact that the raters in the present study were able to distinguish between the criteria successfully in spite of their inexperience in rating and lack of extended training seem to provide further evidence for the advantages of empirically developed scales over intuitively developed ones.

The fact that the infit and outfit mean square values of the criteria were within the acceptable levels suggests that these four criteria (i.e., content, organization, language use and vocabulary) relate to the same general dimension, the writing construct. This is termed as psychometric unidimensionality (Eckes, 2009). The same finding was demonstrated in Park’s (2004) study. In this situation, Park (2004) claims that analytic scores can be meaningfully combined to report a single score. The fact that different traits of the rating scale were all related to the same writing construct may provide further evidence for the validity of the rating scales used in the study.

East (2009) claims that inter-rater and intra-rater measurements of reliability might provide further evidence for scale reliability. The point biserial correlation indices were used to examine inter-rater reliability. The point biserial correlation measures indicated significant levels of inter-rater reliability (i.e., Rater #F =0.90, Rater #T=0.91, Rater #Y= 0.88). For intra-rater reliability, rater fit indices were used. All the fit values were close to 1.0 within the perfectly acceptable levels, providing evidence for the consistency within the raters. In other words, this particular group of raters rank-ordered the students somewhat similarly and consistently, providing additional evidence for the fact that raters applied the rating scales reliably in their ratings.

Knoch (2007) argues that a well discriminating rating scale should result in finer separation of student proficiency. The student separation index as indicated in Table 10 is 9.46, suggesting about nine statistically distinct levels of proficiency among this particular group of students. The number of student proficiency levels identified by FACETS was one level more than the number of categories used in the rating scales to distinguish students’ proficiency, providing evidence for the fact that the rating scales were satisfactorily well-discriminating among the proficiency levels of the students.

Linacre (2002) mentions the usefulness of some category statistics including average student measure, outfit mean- square values and Rasch-Andrich threshold measures in evaluating the effectiveness of rating scale categories. As suggested by Linacre, the average measure by rating scale category and Rasch- Andrich thresholds advanced monotonically as the categories increased from 1 to 8. In other words, “the higher the category [was], the larger the average measure” (Eckes, 2009, p. 26). Besides, the outfit mean-square values were all smaller than 2.0 and close to 1.0 as suggested to be able to claim that rating scale categories function as intended. These findings from category statistics seem to provide empirical evidence for the fact that 8 categories of the rating scales were appropriately ordered and satisfactorily distinguishable. The fact that raters were able to distinguish between 8 categories as identified in the rating scales and used them appropriately may suggest another evidence for validity of the rating scales. However, as Myford (personal e-mail communication, November 8, 2016) cautioned, the rating scales should be used with much larger number of students and raters in order to obtain more accurate and stable values of category statistics  and to be able to make sound claims about how well rating scale categories function.

Research question 3: To what extent is the quality of ratings influenced by rater effects?

Just as the type of rating scale used to rate student responses, raters involved in the ratings of those responses have the potential to contribute to score variability, which is generally conceptualized as rater effects. Therefore, it is important to examine the effects of raters on the quality of ratings to be able to make validity claims about test scores. Previous studies have demonstrated considerable rater effects (Eckes, 2005; Engelhard, 1994; Leckie & Baird, 2011). This study focused on four types of rater effects through many-facet Rasch analysis: rater severity, central tendency, halo effect and inconsistency.

The MFRM analysis provided useful information to detect and evaluate rater effects that might have been involved in the ratings of student responses. Selected statistics from the rater measurement report suggested that raters differed somewhat in their severity with which they rated student responses. However, the difference between the most severe and the most lenient rater was not substantial (0.33 logit). In other words, the raters exercised similar levels of severity although they were not interchangeable. Rater fit indices indicated the raters were consistent in the way they applied the scoring criteria. These findings may provide further evidence for the findings of previous research (i.e., Weigle, 1998) about the effect of rater training. It is considered that rater training is effective for improving raters' self-consistency rather than eliminating the differences in rater severity. That is to say, rater training contributes to intra-rater reliability more than inter-rater reliability. However, this is not seen as an ineffectiveness of the rating training in the literature. Lumley and McNamara (1995) argue that the main concern of the rater training should be to minimize "the random error in rater judgments" due to the fact that a lack of self consistency in the ratings makes it impossible to carry out an orderly process of measurement (p.57). Similarly, McNamara (1996) stated,

> To accept that the most appropriate aim of rater training is to make raters internally consistent so as to make statistical modeling of their characteristics possible, but beyond this to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way. (p. 127)

These statements imply the priority of intra-rater consistency over the inter-rater reliability in evaluating the effectiveness of rater training. However, the point biserial correlation measures in Table 14 also indicated a significant level of reliability among the raters in the current study although some levels of severity difference existed. Therefore, the raters in the current study seemed to have applied the rating scales consistently and similarly, and a lack of inter-rater reliability was not a problem in the study either.

Central tendency is observed when the raters tend to overuse middle scale categories or avoid using extreme categories (Leckie & Baird, 2011). Evidence for the absence of central tendency effect was obtained from fit statistics and the measure of student separation index as suggested by Myford and Wolfe (2004). The fact that there were no overfitting raters suggested that raters did not tend to overuse certain categories. The tendency of overusing certain categories is generally observed for middle scale categories, and raters appear to be too consistent in such cases. Besides, the fact that the ratings for the two tasks reliably separated students into nine distinct levels of proficiency (i.e., student separation index = 9.46) is another indication of the absence of central tendency. This finding suggests that raters made use of all the scale categories in their ratings and they were able to distinguish between the performances of students that displayed different levels of proficiency.

Halo effect occurs when raters fail to differentiate conceptually distinct aspects of a rating scale, and tend to score student performance holistically (Engelhard, 1994). The fact that halo effect did not appear with this group of raters was already discussed in the second question to investigate the reliability of the rating scales. The criteria separation index suggested that raters applied the rating scales analytically as intended.

When the four types of rater effects examined, it seems that there is not much evidence for rater-related variance. One possible explanation for these findings might be the overall effectiveness of the rater training despite the fact that it was brief. It is often stressed in literature that having a well-developed rating scale with clear and explicit level descriptors would be insufficient without exemplifying those level descriptors with actual student responses through rater training (i.e., Shaw & Weir, 2007). Along with the effectiveness of rater training, the findings seem to provide evidence for the effectiveness of the empirically developed rating scales. Knoch

(2007) argues that intuitively developed rating scales have the potential to cause various rater effects. As the descriptors of such scales are not explicit enough, raters tend to create their individual interpretations of the descriptors, which in turn may cause rater severity, inconsistencies, halo effect or central tendency if raters simply choose ‘the play-it safe method’ and use the middle categories. Knoch (2007) further argues that empirically-developed rating scales tend to eliminate these types of rater-related variables. The current study seems to provide evidence to claim that empirically developed rating scales can be used appropriately and reliably by the raters. All in all, the fact that serious rater effects were not involved in the ratings provides evidence for the score validity as variances associated with raters are accepted as construct-irrelevant variance and adversely affect score validity.

Research question 4: How dependable are the scores assigned to the students?

The last question examined the reliability of scores using the G- theory analysis. Specifically, the G-theory analysis investigated the relative contribution of students, tasks and raters to the overall variability in the ratings. The findings from G-theory analysis were consistent with the findings of the MFRM analysis, both of which are often used to examine the possible variances involved in the scores obtained from performance tests.

The results from G-studies indicated that the largest source of score variance was explained by the true differences among students’ writing proficiencies.  A large variance that is associated with students is not accepted as an error variance and it is actually desirable.  Among the analytical scores, the scores on language use and vocabulary seem to be the most dependable as 80% and 79% of the total variance was explained by the students’ real differences in their writing proficiencies on these criteria respectively. In other words, the scores assigned for language use and vocabulary were more generalizable than the scores assigned for content and organization (62.3% and 68.1% respectively). This suggests that students were somewhat better discriminated on their scores associated with language use and vocabulary, in which relatively smaller error variance involved. This tells us that although content and organization are important criteria in assessment of writing, the essentiality of grammatical accuracy and lexical adequacy is indispensable.

The results also showed relatively large variance associated with tasks and person-by-task interaction. Relatively large variance explained by the tasks suggested that the tasks differed in terms of difficulty, and this difference was most salient for content. As discussed in question 1, the tasks differed from each other in terms of task characteristics such as input stimuli, discourse mode, cognitive demands, linguistic demands and expected response length, and thus were purposefully expected to be of different difficulty. Therefore, to elicit different levels of performance from the students on each task was the purpose of the study and this finding suggested that this purpose was achieved, however in the opposite direction. Contrary to what was expected, students performed better on the essay task, which was assumed to be cognitively more demanding than the graph interpretation task. In the present study, task-related variance was therefore associated with true variance although difference in task difficulty is generally accepted as error variance in G theory when the two tasks are expected to function similarly. The fact that the tasks in the present study differed in difficulty mostly in terms of content may provide further evidence to suggest that task unfamiliarity is a strong factor in graph interpretation tasks, and students can be disadvantaged considerably. Therefore, students might need special training to be familiarized with the content requirements of the graph interpretation task as it was also revealed in the MFRM analysis and descriptive statistics that the graph interpretation task was more difficult than the essay task for the students to receive high scores in. Person-by-task interaction accounted for about 18% of total variance for content, 14.2% of total variance for organization, 8.8% of total variance for language use and 3.4% of total variance for vocabulary. This finding suggested that the rank-ordering of a number of students was not consistent across the two tasks, and this was the most obvious for content and organization. This finding was consistent with the findings of previous studies (i.e., Gebril, 2009; Lee & Kantor, 2005; Schoonen, 2005; Sudweeks et al, 2005). Gebril (2009) argues that it is not uncommon to find large person-by-task interaction in performance assessment and particularly in writing assessments. However, Sudweeks et al. (2005) claim that "[in such cases] any generalizations about students' relative standing based on either one of the [tasks] by itself would not be dependable and would lead to different conclusions about students' writing abilities" (p. 249). Therefore, it is suggested that students' performance on both tasks should be considered before making any judgments or decisions about students' proficiency or giving decisions.

As opposed to relatively large effect of task component, ratings and the interaction components of person-by-ratings, and task-by-ratings explained a very small amount of score variance, suggesting that the students were rank-ordered somewhat similarly by the first and second ratings. In other words, ratings facet did not significantly contribute to the score variability. These findings confirmed the findings of the MFRM, according to which no substantial rater effects were identified.  As discussed before, small effects of ratings may provide evidence for the effectiveness of rater training in familiarizing the raters with the scoring scales, which was also suggested by Gebril (2009) upon obtaining similar results. This finding also gives substantial support to scoring validity.

A second stage of the question investigated the optimal number of tasks and raters in order to maximize score dependability with the help of D-studies. Overall, the findings from D-studies suggest that increasing the number of tasks may have more considerable impact on score dependability than increasing the number of ratings. As noted by Lee and Kantor (2005), relatively large impact of tasks on score reliability is not unexpected considering the large amount of task-related variances found in G-studies. To illustrate, increasing the number of tasks from 1 to 3 in a single rating scenario, there were dramatic increases in G-coefficients of analytic scores (i.e., 0.18, 0.16, 0.11, and 0.10 increases for content, organization, language use and vocabulary respectively). When the number of tasks was changed from 3 to 4; the increase in G- coefficients was not that much dramatic. Therefore, an optimum number of tasks used can definitely not be “one” in this study. One explanation for this might be the fact the tasks used in the study are not of equal difficulty, and they are two different task types tapping into different constructs of academic writing, or different aspects of the same construct. For this reason, making generalizations about students proficiency based on only one task might be misleading.

As relatively small amount of rating-related variances were observed in the G- studies, it seems that increasing the number of ratings did not lead to dramatic increases in G-coefficients as much as increasing the number of tasks. For example, the G coefficient for content increased from 0.68 to 0.73 in a single task double rating scenario. When the number of ratings further increased to 3 in a single task scenario, there was only a 0.1 gain in the G coefficient. Among the various D-study scenarios, three tasks with double ratings seem to lead to significantly high levels of reliability in scores (i.e., 0.89 for content, 0.91 for organization, 0.95 for language use and 0.96 for vocabulary) in this study. However, considering the satisfactory values of G-coefficients for 2 tasks and double ratings (i.e., 0.84 for content, 0.87 for organization, 0.93 for language use and 0.95 for vocabulary), it seems reasonable to argue that 2 tasks and double ratings used for the study were able to generate reliable scores, and thus can be used in this way for further studies. However, if the practicality is not a primary concern, three tasks may also be used in order to obtain more dependable and generalizable scores.

To summarize, the present study provided evidence from various perspectives to support validity of the newly developed TFL writing test, which was the main consideration of the study.  In order to gather such evidence, attempts were made to find empirical evidence for scoring validity through investigating the effectiveness of the rating scales used, the possible involvement of rater effects and dependability of the test scores. In addition, the differing cognitive and linguistic demands of the two tasks were discussed and their effect on students' performance was investigated through a posteriori analysis of the test scores to provide evidence for cognitive and contextual validity. The findings demonstrated successful implementation of the scoring procedures that resulted in high scoring validity, and finely-discriminated proficiency levels elicited from the two different tasks, which also supported cognitive validity of the tasks.

# CHAPTER 6

# CONCLUSION

## 6.1 Introduction

The present study was conducted to investigate and provide validity evidence for the writing proficiency test developed for students of TFL at Boğaziçi University. The test consists of two tasks (the graph interpretation task and the essay task) that are characterized with different cognitive demands and contextual features. The tasks were intended to be at different proficiency levels proposed in the CEFR (B2 and C1). They were examined in terms of cognitive, context and scoring validity aspects of the validity framework proposed by Weir (2005). The main findings of the study were as follows:

First, the two writing tasks that differed in terms of their intended CEFR levels, cognitive demands and contextual features led to differences in students' performance although the difference was in the opposite direction of the expectation of the researcher. It was hypothesized that students would receive higher scores on the graph interpretation task as the graph interpretation task was assumed to be cognitively less demanding in terms of length, language and organizational requirements, and thus easier for students to get high scores in. However, it was found out that students performed better in the essay task, which was considered to be cognitively more demanding, on all the criteria of the scoring scale (i.e., content, organization, grammar and vocabulary). This finding raised the issue of task unfamiliarity that is often associated with graph-writing tasks (Yang, 2012), and the need for special training to familiarize students with such type of tasks.

Second, although students performance did not correspond to the expectation of the researcher, both of the tasks were able to discriminate well between different proficiency levels of students. This provided evidence for the fact that both of the tasks functioned well as intended.

Third, it was found that the scoring scales developed through empirically-based method by analyzing students' responses were applied consistently and appropriately by the raters. The fact that serious rater effects were not involved in the ratings of the student responses despite the inexperience of raters and the lack of extended training provided evidence for scale effectiveness and the clarity of wording of the scale descriptors. The preference of analytical scale proved to be useful in pinpointing the uneven profile of learners of Turkish. For example, it was found that students performed best on vocabulary while they performed worst on content aspect.

Fourth, the findings indicated that Many-facet Rasch measurement is a useful measurement model in identifying task difficulty, scale effectiveness and raters effects, which are potential sources of variance that can affect validity of performance tests.

Finally, G-theory analysis revealed that the scores assigned to students by the two raters were highly dependable (reliable) based on the fact that the largest percentage of variance was explained by the actual differences in the students' proficiency, and raters had a very little effect on the score variance. D-studies indicated that increasing the number of tasks had a more dramatic impact on score dependability than increasing the number of ratings. However, because of satisfactorily reliable test scores that were obtained by the current study with 2 tasks and 2 raters scenario, no further suggestions are made on the number of tasks and ratings.

To conclude, the findings seem to provide certain evidence for the validity of the writing test of TFL in terms of cognitive, contextual and scoring validity aspects. However, it should be noted that the conclusions drawn here are preliminary and tentative based on a small number of student responses and a small group of raters. In order to make stronger validity claims, further research with larger number of participants and raters is needed.

## 6.2 Research implications

Based on the findings of the present study, it can be argued that different types of writing tasks should be involved in a test in order to obtain a more accurate picture of test takers' performance as it is clear that the task type used tends to affect test taker performance. However, tasks that test takers are not quite familiar with such as a graph interpretation task should be used with caution as such tasks may introduce construct-irrelevant variances in performance  due to graph unfamiliarity and a lack of task-related vocabulary. What is more, special training on this type of graph interpretation tasks may be necessary in order to minimize these kinds construct irrelevant variances.

Second, the study suggests that rating scales that are developed through empirically-based method leads to high levels of inter-rater and intra-rater reliability as the raters are able to use such scales consistently and appropriately. Therefore, it can be claimed that institutions should make use of empirically-based scoring scales in the assessment of writing proficiency although intuitive-based scales are easier to construct, and thus used more frequently. Moreover, the findings suggest that analytical scales are able to pinpoint uneven aspects of students' performance as it may often be the case with second language learners and they better reflect the aspects of performance that might be affected by such construct-irrelevant variables as task unfamiliarity. Therefore, it is suggested that in the assessment of second language learners' writing ability, analytical scoring scales can be preferred over holistic scales even though holistic scoring scales are more often preferred due to practicality reasons.

Third, the findings prove that many-facet Rash measurement can be successfully used in identifying factors that have the potential to influence validity of performance tests such as task difficulty, criteria difficulty, rating scale reliability and possible rater effects. Thus, it is suggested that many-facet Rasch measurement model should be employed by the researchers who are involved in test validation studies as well as institutions who deliver subjectively scored writing tests.

### 6.3 Limitations and suggestions for further research

One of the major limitations of the study was the small sample size as noted a few times before. Because of the small number of data, the conclusions drawn in this study are tentative and preliminary. With a much larger number of participants and raters, the analysis of many-facet Rasch measurement and generalizability theory could have produced more reliable results, and sounder claims about test validity and rating scale effectiveness could have been made.

Another limitation that should be mentioned resulted from the restrictions on the width of the study. Even though the procedures that were followed during the task and rating scale development constituted a large part of the study, and they are valuable in evaluating the validity of the test, it was not possible to report the whole process and include it in the discussion of the study because of space limitations.

Third, the effect of task types on students' performance was investigated only based on the test scores. For future research, a thorough analysis of syntactic variation (i.e., the length and accuracy of T-units, vocabulary range) of students' responses might be more illuminating to draw more accurate conclusions. It will definitely be worthwhile to compare the cognitive processes used by the test takers in graph interpretation and essay tasks through think aloud procedures in order to get a more in-depth understanding of the requirements of them.

Lastly, the present study examined validity in terms of cognitive, context and scoring validity aspects while consequential validity and criterion validity were beyond the scope of the study. Comparing the results of the study with the results of a high stakes test such as IELTS, for example, would result in further evidence for validity claims. Therefore, these aspects of validity may be the focus of another study.

All in all, although this study has provided substantial evidence as to the accuracy of possible decisions to be made based on this test's results, it has also shown that test validation is an ongoing and a comprehensive process. The more evidence is collected, the sounder the validity claims about the usefulness and appropriateness of the tests are.

# APPENDIX A

## TASK SPECIFICATIONS

General Description (Final Draft)

The writing test of Turkish aims to assess candidates' academic writing ability in two levels as proposed in the CEFR:  B2 and C1. The difference between the levels is determined according to type of text to be produced, and the cognitive and linguistic demands of tasks.  At B2 level, students need to produce clear and detailed texts on a variety of subjects within their interests by using a wide range of structure and vocabulary. At C1 level, the candidates are expected to write detailed and well-structured texts on complex subjects using complex structures and a more precise and more varied vocabulary. The test content is informed by the CEFR level descriptors. The test consists of two tasks, one of which is characterized as an integrated task and the other is an independent task. The topics are of general interest and lead to no privilege among candidates studying different subjects. Candidates are assessed on their ability to how effectively they fulfill organizational, linguistic, and content requirements of the given task type.

Task 1

| *Task Input/Prompt* | |
|---|---|
| CEFR level | B2 |
| Task content | CEFR B2: Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources. Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail.<br>Can evaluate different ideas or solutions to a problem. |
| Time permitted or suggested for this task | 20 minutes |
| Content | Fully-specified |
| Task type (genre) | Report |

| | |
|---|---|
| Format | Candidates are required to deal with visual data in the form of charts or graphs and input of up to 50 words. |
| Task focus | Describing, summarizing, interpreting, comparing |
| Mode of input | Written and visual |
| Theme of input | Education |
| Integration of skills | Visual data |
| *Response* | |
| Number of words expected | 100-150 |
| Rhetorical functions expected | describing, summarizing, interpreting, comparing |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge telling, graph comprehension, graph interpretation and graph translation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

Task 2

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | C1 |
| Task content | CEFR C1: Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. Can write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. |
| Time permitted or suggested for this task | 30 minutes |
| Content | Specified to some extent |
| Task type (genre) | Essay |

| | |
|---|---|
| Format | Candidates are presented with a point of view or an argument with an input of up to 40 words. |
| Task focus | Presenting and justifying an opinion, evaluating and challenging ideas, explaining advantages and disadvantages of various options |
| Mode of input | Written |
| Theme of input | Social Media |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 250-300 |
| Rhetorical functions expected | Presenting and justifying an opinion, evaluating and challenging ideas, explaining advantages and disadvantages of various options. |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge transformation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

### Initial Task Specifications

### General Description

The writing test aims to assess candidates’ academic writing ability in three levels proposed in the CEFR which are B1, B2 and C1. The difference between the levels is determined according to the type of text to be produced, the complexity of the topics, and the cognitive and linguistic demands of tasks. At B1 level, the candidates are supposed to write relatively short and simple connected text using relatively simple structures and cohesive devices on familiar topics whereas at B2 level, they need to produce clear and detailed texts on a variety of subjects within their interests by using a wider range of structure and more varied vocabulary. Finally, at C1 level, the candidates are expected to write detailed and well-structured texts on complex subjects using complex structures and a more precise and more varied vocabulary. At all levels, the Writing paper consists of two or three tasks. The topics are of general interest and lead to no privilege among candidates studying different subjects.

### Test Focus

According to Common European Framework of Reference,

At B1 level candidates are assessed on their ability to:

- write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.
- write accounts of experiences, describing feelings and reactions in simple connected text.
- write a description of an event, a recent trip – real or imagined.
- narrate a story.
- write short, simple essays on topics of interest.
- summarize, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his/her field with some confidence.
- write very brief reports to a standard conventionalized format, which pass on routine factual information and state reasons for actions.

At B2 level candidates are assessed on their ability to:

- write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources.
- write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned.
- write a review of a film, book or play.
- write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail.
- evaluate different ideas or solutions to a problem.
- write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
- synthesize information and arguments from a number of sources.

At C1 level candidates are assessed on their ability to:

- write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.
- write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind.
- write clear, well-structured expositions of complex subjects, underlining the relevant salient issues.

TASK 1

| *Task Input/Prompt* | |
|---|---|
| CEFR level | B1 |
| Time permitted or suggested for this task | 40 minutes |
| Content | Fully-specified |
| Task type (genre) | An email |
| Format | Candidates are required to respond to a given situation specified in no more than 50 words. |

| | |
|---|---|
| Task focus | Describing, explaining, commentating, suggesting, warning |
| Mode of input | Written |
| Theme of input | Places |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 150-200 |
| Rhetorical functions expected | Describing (places), explaining, commentating, suggesting, warning |
| Register | Informal |
| Domain | Personal |
| Cognitive processing | Knowledge telling |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK2

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | B1 |
| Time permitted or suggested for this task | 40 minutes |
| Content | Fully-specified |
| Task type (genre) | story |
| Format | A situation-based writing task specified in no more than 40 words. |
| Task focus | Narrating, reporting events, describing experiences and feelings |
| Mode of input | Written |
| Theme of input | Memories |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 150-200 |

| | |
|---|---|
| Rhetorical functions expected | Narrating, reporting events, describing experiences and feelings |
| Register | Neutral |
| Domain | Personal |
| Cognitive processing | Knowledge telling |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK 3

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | B2 |
| Time permitted or suggested for this task | 30 minutes |
| Content | Fully-specified |
| Task type (genre) | Report |
| Format | Candidates are required to deal with visual data in the form of charts or graphs and input of up to 40 words. |
| Task focus | Describing, summarizing, interpreting, comparing |
| Mode of input | Written and visual |
| Theme of input | Technology |
| Integration of skills | Visual data |
| *Response* | |
| Number of words expected | 100-150 |
| Rhetorical functions expected | Describing, summarizing, interpreting, comparing |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge telling, graph comprehension, graph interpretation and graph translation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |

| | |
|---|---|
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK 4

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | B2 |
| Time permitted or suggested for this task | 30 minutes |
| Content | Fully-specified |
| Task type (genre) | Review |
| Format | A situation-based writing task specified in no more than 50 words. |
| Task focus | Describing, expressing opinion, recommending |
| Mode of input | Written |
| Theme of input | Films |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 150-200 |
| Rhetorical functions expected | Describing, expressing opinion, recommending |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge telling |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK 5

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | B2 |
| Time permitted or suggested for this task | 45 minutes |
| Content | Fully-specified |

| | |
|---|---|
| Task type (genre) | Essay |
| Format | Candidates are presented with a problem given with an input of up to 50 words. |
| Task focus | Evaluating a problem, presenting solutions to a problem. |
| Mode of input | Written |
| Theme of input | Education |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 200-250 |
| Rhetorical functions expected | Describing, expressing opinion, recommending |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge transformation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK 6

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | C1 |
| Time permitted or suggested for this task | 45 minutes |
| Content | Fully-specified |
| Task type (genre) | Essay |
| Format | Candidates are presented with a point of view or an argument with an input of up to 40 words. |
| Task focus | Presenting and justifying an opinion, comparing and contrasting, evaluating and challenging ideas or arguments. |
| Mode of input | Written |
| Theme of input | Life style |
| Integration of skills | None |
| *Response* | |

| | |
|---|---|
| Number of words expected | 250-300 |
| Rhetorical functions expected | Presenting and justifying an opinion, comparing and contrasting, evaluating and challenging ideas or arguments |
| Register | Formal |
| Domain | Academic |
| Cognitive processing | Knowledge transformation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

TASK 7

| | |
|---|---|
| *Task Input/Prompt* | |
| CEFR level | C1 |
| Time permitted or suggested for this task | 45 minutes |
| Content | Fully-specified |
| Task type (genre) | Essay |
| Format | Candidates are presented with a point of view or an argument with an input of up to 40 words. |
| Task focus | Presenting and justifying an opinion, evaluating and challenging ideas, explaining advantages and disadvantages of various options |
| Mode of input | Written |
| Theme of input | Social Media |
| Integration of skills | None |
| *Response* | |
| Number of words expected | 250-300 |
| Rhetorical functions expected | Presenting and justifying an opinion, evaluating and challenging ideas, explaining advantages and disadvantages of various options. |
| Register | Formal |
| Domain | Academic |

| | |
|---|---|
| Cognitive processing | Knowledge transformation |
| Content knowledge required | General/non specialized |
| *Rating of Task* | |
| Rating method | Analytic |
| Assessment criteria | Content, organization, language use, vocabulary |
| Number of raters | 2 |

# APPENDIX B

## WRITING TASKS

*Writing Tasks (Final Draft)*

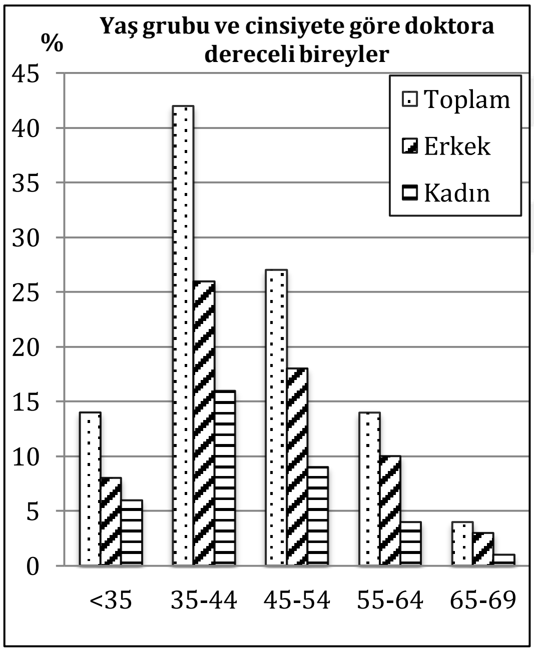
**Sınav 1**

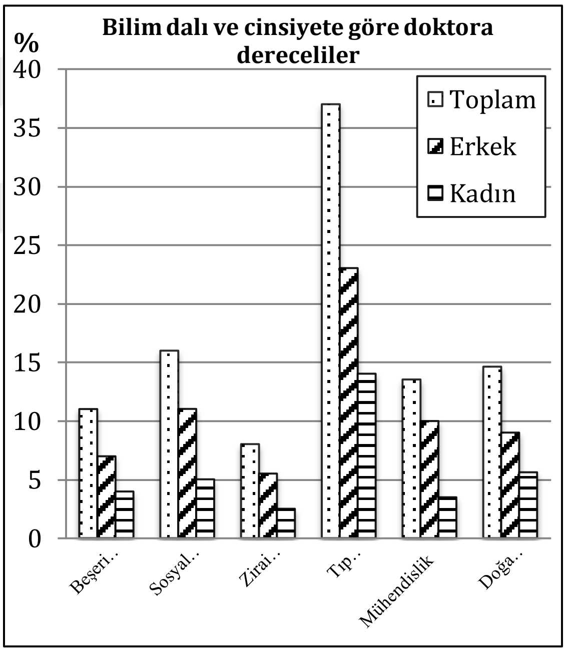
**Aşağıda Türkiye'de *doktora derecesine sahip bireylerin 2008 yılı istatistikleri* verilmiştir. İlk grafik yaş ve cinsiyete göre, ikinci grafik bilim dalı ve cinsiyete göre doktora derecesine sahip bireylerin yüzdelerini göstermektedir.**

**Yönerge:** Bu iki grafikteki bilgileri kullanarak bir rapor yazınız. Raporunuzda grafiklerdeki önemli bilgileri belirtiniz ve gerektiğinde karşılaştırma yapınız.

Bu bölümü tamamlamak için en az 150 sözcük kullanınız.

Süreniz 20 dakikadır.





Source: Türkiye İstatistik Kurumu (TÜİK)

**Sınav 2**

**Yönerge:** Aşağıdaki konuyla ilgili bir deneme yazınız. Bu bölümü tamamlamak için en az 250 sözcük kullanınız. Süreniz 30 dakikadır.

**“Günümüzde yaygın olarak kullanılan *Facebook, Twitter* gibi sosyal medya sitelerinin insanlara verdiği zararlar sağladığı faydalardan daha fazladır.”**

**Siz bu konuda ne düşünüyorsunuz? Konuyla ilgili kendi görüşlerinizi nedenleriyle açıklayınız ve örneklerle destekleyiniz.**

*Writing Tasks (Used in the Second Phase)*

**Sınav 1.** Başınızdan geçen ilginç bir olayı anlatınız. Bu bölümü tamamlamak için en az 180 kelime kullanınız. Süreniz 30 dakikadır.

**Yazınızda şunlardan söz ediniz:**

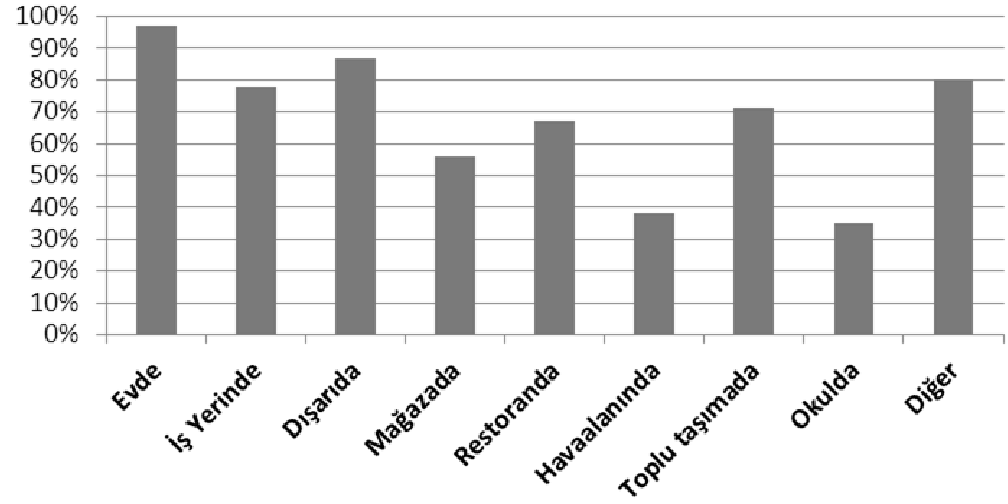
- **Olayın geçtiği yer ve zaman**
- **Olayla ilgili kişiler**
- **Önemli olaylar**
- **Hikayenin sonu**

**Sınav 2**

Aşağıda verilen grafiklerdeki bilgilere dayanarak Türkiye'de *akıllı telefon* kullanımı konusunda bir rapor yazınız. Bu bölümü tamamlamak için en az 150 kelime kullanınız.

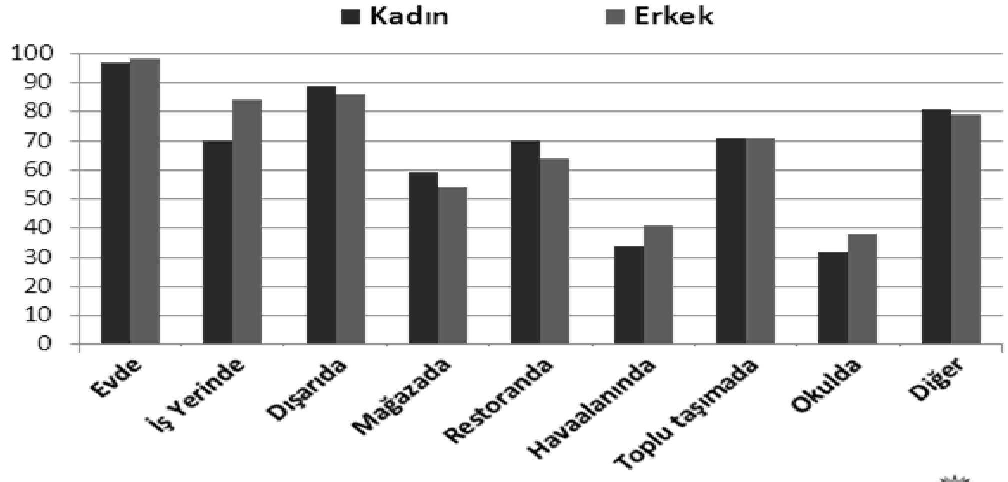
Süreniz 20 dakikadır.

**Grafik 1: Türkiye genelinde akıllı telefonların kullanıldığı yerler**



Kaynak: Our Mobile Planet, Mart-Temmuz 2011

**Grafik 2: Türkiye'de cinsiyete göre akıllı telefonların kullanıldığı yerler**



Kaynak: Our Mobile Planet, Mart-Temmuz 2011

**Sınav 3**

**Yönerge:** Aşağıdaki konuyla ilgili bir deneme yazınız. Bu bölümü tamamlamak için en az 250 kelime kullanınız. Süreniz 40 dakikadır.

**"Günümüzde yaygın olarak kullanılan *Facebook, Twitter* gibi sosyal medya sitelerinin insanlara verdiği zararlar sağladığı faydalardan daha fazladır."**

**Siz bu konuda ne düşünüyorsunuz? Konuyla ilgili kendi görüşlerinizi nedenleriyle açıklayınız ve örneklerle destekleyiniz.**

*Writing Tasks (Initial Draft)*

**Sınav 1**

**Yönerge:** Sınavın bu bölümünde aşağıda verilen konuyla ilgili bir metin yazınız. Bu bölümü tamamlamak için en az 180 kelime kullanınız. Süreniz 40 dakikadır.

**Farklı ülkeden bir arkadaşınız sizin yaşadığınız şehri ziyarete gelmek istiyor. Arkadaşınıza yardımcı olmak için bir e-mail yazarak şehrinizi tanıtınız.**
**E-mailinizi yazarken aşağıdaki noktalara değinmelisiniz.**

- **Şehrin coğrafi/fiziki özellikleri (konumu, nüfusu, ulaşım vb.)**
- **Gezilip görülecek yerler**
- **Yapılacak aktiviteler, sosyal yaşam**
- **Uyarılar, tavsiyeler**

**Sınav 2**

**Yönerge**: Sınavın bu bölümünde aşağıda verilen duruma göre bir metin yazınız. Bu bölümü tamamlamak için en az 180 kelime kullanınız. Süreniz 40 dakikadır.

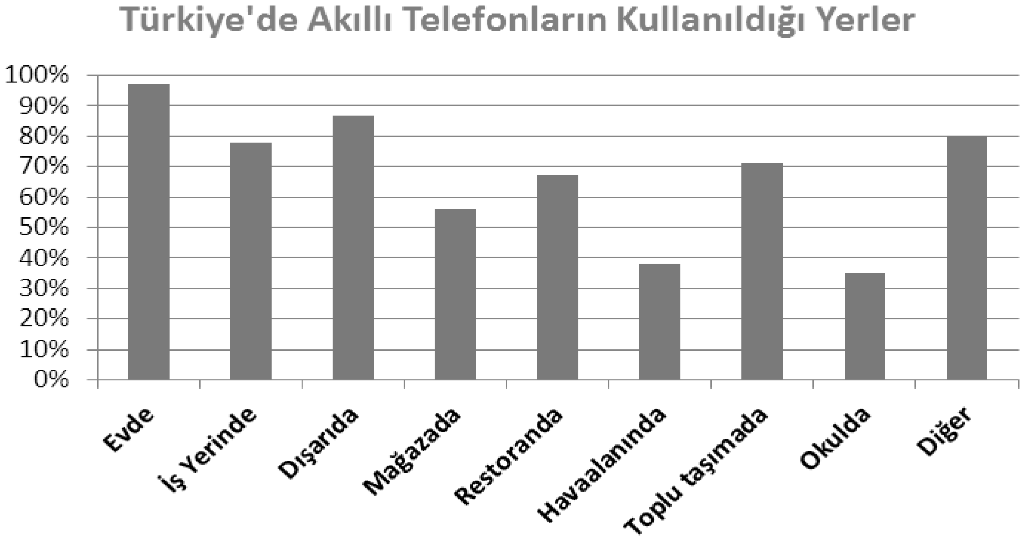
**Edebiyat bölümü hocanız öğrencilerin ödevleri için bir blog oluşturdu. Ödev olarak, başınızdan geçen ilginç bir olayı anlatan yazı yazmanızı istedi. Hocanız, en ilginç hikâyeyi seçip "Haftanın Yazısı" köşesine ekleyecek. Yazacağınız hikâyede şunlara dikkat etmelisiniz.**

- **Hikâyenizi en az üç paragrafa ayırınız: giriş, gelişme, sonuç**
- **Olayın mekânını, zamanını belirtiniz ve olaydaki kişiler hakkında bilgi veriniz.**
- **Olayları kronolojik sıraya göre anlatınız.**

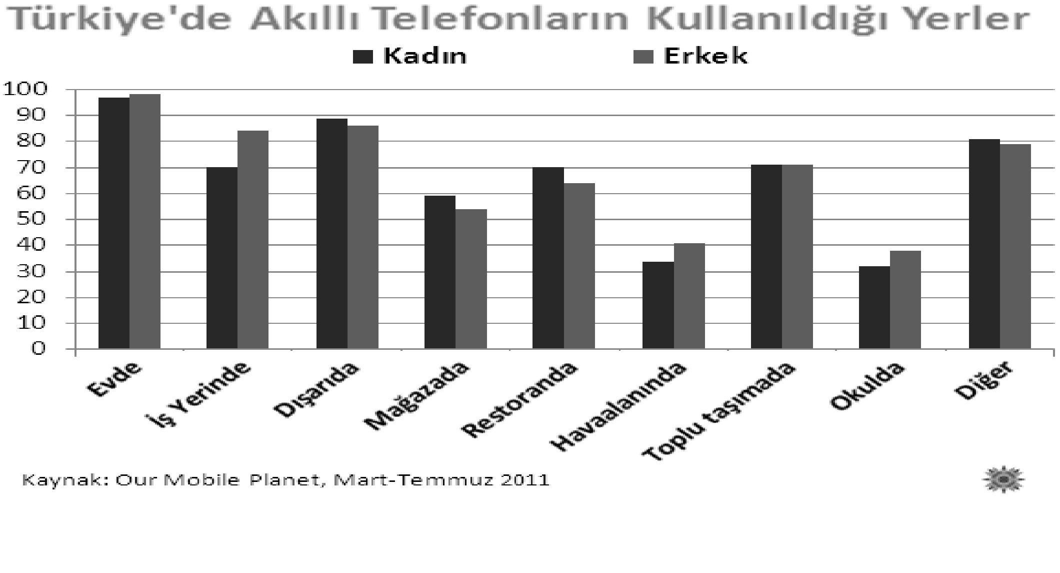
**Sınav 3**

**Yönerge:** Sınavın bu bölümünde aşağıda verilen grafiklerdeki bilgileri kullanarak bir rapor yazınız. Bu bölümü tamamlamak için en az 180 kelime kullanınız. Süreniz 30 dakikadır.

> 1. **Grafik, Türkiye'de akıllı telefon kullanıcılarının mobil cihazlarını en çok nerelerde kullandığını gösteriyor.**
> 2. **Grafik, akıllı telefon kullanıcılarının cinsiyete göre (kadın-erkek) en çok nerelerde kullandığını gösteriyor.**
>
> **Bu iki grafikte verilen bilgileri özetleyip yorumlayınız. Bu bilgileri sentezleyerek ve karşılaştırma yaparak rapor ediniz.**



Kaynak: Our Mobile Planet, Mart-Temmuz 2011



Kaynak: Our Mobile Planet, Mart-Temmuz 2011

**Sınav 4**

**Yönerge:** Sınavın bu bölümünde aşağıda verilen duruma göre bir metin yazınız. Bu bölümü tamamlamak için en az 200 kelime kullanınız. Süreniz 30 dakikadır.

**Kitap, tiyatro ve film eleştirileri yazan bir kültür dergisinde çalışıyorsunuz. Derginin editörü sizden son zamanlarda izlediğiniz bir filmin eleştirisini yazmanızı istedi. Yazacağınız film eleştirisinde şu noktaları ele alınız:**

- **Filmin genel özellikleri (türü, yılı, süresi, oyuncular, yönetmen, vb.)**
- **Kısaca hikâyenin konusu**
- **Filmin beğendiğiniz ve beğenmediğiniz yönleri**
- **Okuyucuya tavsiyeler**

**Sınav 5**

**Yönerge**: Sınavın bu bölümünde aşağıda verilen konuya ilişkin kendi görüşlerinizi belirten bir metin yazınız. Bu bölümü tamamlamak için en az 250 kelime kullanınız. Süreniz 45 dakikadır.

**Uzmanlara göre "sınav kaygısı" birçok öğrencinin başarısız olmasının sebeplerinden biridir. Bununla birlikte, bazı önlemler alarak bu problemi aşmak mümkündür.**

**Sizce sınav kaygısı üstesinden gelmek için neler yapılabilir? Metninizi yazarken şunlara dikkat ediniz:**

- **Problemi ve neden çözülmesi gerektiğini belirtiniz.**
- **Çözümleri ve bu çözümlerden beklediğiniz sonuçları belirtiniz.**
- **Düşüncelerinizi nedenleriyle açıklayınız ve örneklerle destekleyiniz.**

**Sınav 6**

**Yönerge:** Sınavın bu bölümünde aşağıda verilen konuya ilişkin kendi görüşlerinizi belirten bir deneme yazınız. Bu bölümü tamamlamak için en az 250 kelime kullanınız. Süreniz 45 dakikadır.

**Bazı insanlar kırsal kesimlerde çocuk yetiştirmenin şehir merkezinde çocuk yetiştirmekten daha doğru olduğunu iddia ediyor. Siz bu görüşe katılıyor musunuz? Metninizi yazarken şunlara dikkat ediniz:**

- **Düşüncelerinizi gerekçelendirerek ve örnekler vererek destekleyiniz.**
- **Her iki seçeneğin avantajlarını ve dezavantajlarını karşılaştırınız. (Kır hayatı-şehir hayatı)**
- **Kendi tercihinizi sebepleriyle birlikte belirtiniz.**

**Sınav 7**

**Yönerge:** Sınavın bu bölümünde aşağıdaki konuyla ilgili bir deneme yazınız.
Bu bölümü tamamlamak için en az 250 kelime kullanınız. Süreniz 45 dakikadır.

**Bir televizyon programında şu konunun tartışıldığını gördünüz:**
**"Günümüzde yaygın olarak kullanılan Facebook, Twitter gibi sosyal medya sitelerinin insanlara sağladığı faydaların yanı sıra bir takım zararları da olmuştur."**
**Siz bu konuda ne düşünüyorsunuz? Metninizi yazarken şunlara dikkat ediniz:**

- **Konuyla ilgili kendi görüşlerinizi nedenleriyle açıklayınız ve örneklerle destekleyiniz.**
- **Okuyucuyu ikna etmek için, karşıt görüşleri de ele alıp kendi görüşünüzle çürütünüz.**

# APPENDIX C

## SCORING RUBRIC

*Final Draft*

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE-TASK 1

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 8-7 | Fully addresses all the requirements of the task:<br>Clearly and effectively identifies main trends, describes and/or compares important information. |
| | GOOD TO AVERAGE | 6-5 | Adequately addresses the requirements of the task:<br>Presents an overview of main trends, but may skip some important information or may include a few unnecessary details. |
| | FAIR TO WEAK | 4-3 | Attempts to address the task but doesn't cover all the requirements of the task:<br>Unable to give a clear overview of main trends and includes unnecessary details or inaccurate information, or skips important data. |
| | POOR | 2-1 | Fails to address the task or covers just a few points.<br>The response is barely related or completely unrelated to the task. |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 8-7 | Information is well organized and logically sequenced with an order of significance.<br>Information is presented, compared or contrasted meaningfully by using appropriate transition words and phrases. |
| | GOOD TO AVERAGE | 6-5 | Information is generally organized and logically sequenced with an order of significance although important and less important data might be confused at some points.<br>Information is generally connected and the comparison is often meaningful with a range of transition words and phrases, but there may be some over-use/under-use or inaccuracy. |
| | FAIR TO WEAK | 4-3 | Information is presented without a clear logical sequence and progression.<br>Information is generally disconnected and the comparison is often ineffective due to the limited, inaccurate or inappropriate use of transition words and phrases. |
| | POOR | 2-1 | Weak or no organization, sequencing; and no connection and consistency between facts and information. |

| | | | |
|---|---|---|---|
| LANGUAGE USE | EXCELLENT TO VERY GOOD | 8-7 | Uses a wide range of complex structures effectively.<br>Makes only very occasional errors. |
| | GOOD TO AVERAGE | 6-5 | Uses effective but simple structures.<br>Attempts to use complex structures, but these may be less accurate than simple sentences.<br>Makes some errors in grammar and punctuation but they rarely obscure the meaning. |
| | FAIR TO WEAK | 4-3 | Uses a limited range of structures<br>Makes frequent errors in simple and complex structures.<br>Punctuation may often be faulty.<br>These errors frequently impede the meaning. |
| | POOR | 2-1 | Little or no mastery of sentence structures<br>Dominated by errors that blocks the meaning. |

| | | | |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 8-7 | Uses a wide range of task-related vocabulary items with a sophisticated control of lexical features.<br>There may be very occasional errors in spelling and word formation. |
| | GOOD TO AVERAGE | 6-5 | Uses an adequate range of vocabulary items to fulfill the task although there may be some errors in word choice, spelling and word formation.<br>The errors rarely impede the meaning. |
| | FAIR TO WEAK | 4-3 | Uses a limited range of vocabulary items which may be used repetitively or inappropriately.<br>Makes frequent errors in word choice, spelling and word formation.<br>Errors may obscure meaning |
| | POOR | 2-1 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but the meaning is mostly incomprehensible. |

| | |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes a totally memorized response<br>Writes too little (fewer than 30 words) to evaluate. |

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE- TASK2

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 8-7 | Fully addresses all the parts of the task and the topic;<br>Presents a well- developed position to the question with relevant arguments and draw appropriate conclusions.<br>Thorough development of main ideas logically supported with explanations and examples. |
| | GOOD TO AVERAGE | 6-5 | Addresses all parts of the task although some parts may not be elaborated adequately;<br>Presents a position in answer to the question and provide relevant arguments, but the support (elaboration and exemplification) may be limited at certain points. |
| | FAIR TO WEAK | 4-3 | Addresses the task only partially;<br>Unable to express a clear position.<br>Presents some or few arguments with mostly underdeveloped, repetitive or irrelevant support. |
| | POOR | 2-1 | Barely addresses the task;<br>Doesn't express a position,<br>May present a few ideas but there is little or no development<br>Answer may be hardly related or completely unrelated to the task. |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 8-7 | Ideas are well-organized and logically sequenced with beginning, development and ending.<br>Ideas are consistent and connected. Transition between ideas is well managed with effective use of cohesive devices. |
| | GOOD TO AVERAGE | 6-5 | Ideas are generally organized and there is a clear progression in the response.<br>Ideas are generally consistent and connected.<br>Uses a range of cohesive devices, but there may be some over-use/under-use or inaccuracy.<br>May not always use referencing appropriately and clearly. |
| | FAIR TO WEAK | 4-3 | Ideas are presented but not organized logically, and there is no clear overall progression.<br>Uses some basic cohesive devices but these tend to be inaccurate or inappropriate.<br>Ideas may often be disconnected and/or repetitive because of lack of referencing, substitution and linking words. |
| | POOR | 2-1 | Weak or no organization, sequencing; and no connection and consistency between ideas. |

| | | | |
|---|---|---|---|
| LANGUAGE USE | EXCELLENT TO VERY GOOD | 8-7 | Uses a wide range of complex structures effectively.<br>Makes only very occasional errors. |
| | GOOD TO AVERAGE | 6-5 | Uses effective but simple structures. |

|  |  |  |  |
|---|---|---|---|
|  |  |  | Attempts to use complex structures, but these may be less accurate than simple sentences.<br>Makes some errors in grammar and punctuation but they rarely obscure the meaning. |
|  | FAIR TO WEAK | 4-3 | Uses limited range of structures<br>Makes frequent errors in simple and complex structures.<br>Punctuation is often faulty.<br>These errors may frequently impede the meaning. |
|  | POOR | 2-1 | Little or no mastery of sentence structures.<br>Dominated by errors that blocks the meaning. |

|  |  |  |  |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 8-7 | Uses a wide range of vocabulary items with a sophisticated control of lexical features.<br>Effectively uses infrequent lexical items, but there may be occasional errors in word choice and collocations<br>Produces no or rare errors in spelling and word formation. |
|  | GOOD TO AVERAGE | 6-5 | Uses an adequate range of vocabulary items.<br>Attempts to use less frequent vocabulary items, but with some inaccuracy.<br>Makes some errors in spelling and word formation, but they rarely impede the meaning. |
|  | FAIR TO WEAK | 4-3 | Uses a limited range of vocabulary items which may be used repetitively or inappropriately.<br>Makes frequent errors of word formation, word choice and spelling.<br>Errors may often obscure the meaning. |
|  | POOR | 2-1 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but their meanings are incomprehensible in the given context. |

|  |  |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes too little (fewer than 40 words) to evaluate.<br>Writes a totally memorized response. |

*Second Draft (Used in the Second Phase)*

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE-TASK 1 (STORY)

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 7-8 | Fully covers all the content points specified in the task,<br>Adequately expand on the content points through effective descriptions and details. |
| | GOOD TO AVERAGE | 5-6 | Covers the content points although some elements may not be developed adequately or may be omitted.<br>May not elaborate on the content points sufficiently. |
| | FAIR TO POOR | 3-4 | Doesn't cover major content points.<br>Inadequate or little development of the elements of the task.<br>There may be include irrelevant information |
| | POOR | 1-2 | Answer is hardly related or completely unrelated to the task, includes mostly irrelevant information or no development of the content points |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 7-8 | Ideas and events are well-organized and chronologically ordered with beginning, development and ending.<br>Ideas are consistent and connected. Transition and the relationship between ideas are well managed with effective use of cohesive devices (referencing, substitution and linking words.)<br>Effectively manages paragraphing |
| | GOOD TO AVERAGE | 5-6 | Ideas and events are presented with some organization.<br>There is an overall unity and progression.<br>Ideas are generally consistent and connected.<br>The logical sequence and connection of ideas may be affected by limited, inaccurate or over-use of cohesive devices.<br>May not always use referencing appropriately and clearly<br>Uses paragraphing but not always logically. |
| | FAIR TO POOR | 3-4 | Ideas and events are presented but not organized logically. There is little progression in the response.<br>Uses some basic cohesive devices but these are often inaccurate or inappropriate.<br>Ideas are often disconnected, repetitive or mechanical because of lack of referencing, substitution and linking words.<br>May not write in paragraphs |
| | POOR | 1-2 | Weak or no organization, sequencing; and no connection and consistency between ideas. |

| | | | |
|---|---|---|---|
| GRAMMAR | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of complex structures effectively.<br>Makes few errors of subject-verb agreement, tense, number, word order, pronouns , negation , vowel |

| | | | |
|---|---|---|---|
| | | | harmony and case-markers. |
| | GOOD TO AVERAGE | 5-6 | Uses effective but simple structures.<br>Attempts to use complex structures, but minor problems in complex structures.<br>Several errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony but they rarely obscure the meaning. |
| | FAIR TO POOR | 3-4 | Uses limited range of structures<br>Major problems in simple and complex structures.<br>Frequent errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony<br>These errors frequently impede the meaning. |
| | POOR | 1-2 | Little or no mastery of sentence structures<br>Dominated by errors that blocks the meaning. |

| | | | |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of vocabulary with a sophisticated control of lexical features.<br>Effectively uses infrequent lexical items, but there may be occasional errors in word choice and collocations<br>Produces no or rare errors in spelling and word formation. |
| | GOOD TO AVERAGE | 5-6 | Uses an adequate range of vocabulary<br>Attempts to use less frequent vocabulary, but with same inaccuracy.<br>Makes some errors in spelling and word formation, but they do not obscure communication. |
| | FAIR TO POOR | 3-4 | Uses a limited range of vocabulary which may be used repetitively or inappropriately.<br>Makes frequent errors of word formation and spelling<br>Errors may obscure meaning |
| | POOR | 1-2 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but meaning is incomprehensible. |

| | |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes a totally memorized response |

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE-TASK 2 (REPORT)

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 7-8 | Fully addresses all the points specified in the task,<br>Clearly and effectively presents and highlights the main features, describes and/or compares the data, and identifies trends in the given data |
| | GOOD TO AVERAGE | 5-6 | Generally addresses the elements of the task.<br>Presents an overview of main trends or differences but could be more extended.<br>May include irrelevant or inappropriate details. |
| | FAIR TO POOR | 3-4 | Attempts to address the task but doesn't cover all the points specified in the task.<br>May describe the details mechanically with no clear overview of trends.<br>May include unclear, irrelevant, repetitive or inaccurate information. |
| | POOR | 1-2 | Fails to address the task or the task may have been completely misunderstood. The answer is barely related or completely unrelated to the task. |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 7-8 | Ideas and information are well-organized and logically sequenced.<br>Ideas are consistent and connected. Transition between ideas is well managed with effective use of cohesive devices (linking words, referencing, substitution)<br>Effectively manages paragraphing. |
| | GOOD TO AVERAGE | 5-6 | Ideas and information are presented with some organization.<br>There is an overall unity and progression.<br>Ideas are generally consistent and connected.<br>The logical sequence and connection of ideas may be affected by limited, inaccurate or over-use of cohesive devices.<br>May not always use referencing appropriately and clearly.<br>Uses paragraphing but not always logically. |
| | FAIR TO POOR | 3-4 | Ideas and information are presented but not organized logically. There is little progression in the response.<br>Uses some basic cohesive devices but these are often inaccurate or inappropriate.<br>Ideas are often disconnected, repetitive or mechanical because of lack of referencing, substitution and linking words.<br>May not write in paragraphs |
| | POOR | 1-2 | Weak or no organization, sequencing; and no connection and consistency between ideas. |

| | | | |
|---|---|---|---|
| GRAM MAR | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of complex structures<br>Makes few errors of subject-verb agreement, tense, number, word order, pronouns, negation, vowel harmony and case-markers. |

| | | | |
|---|---|---|---|
| | GOOD TO AVERAGE | 5-6 | Uses effective but simple structures.<br>Attempts to use complex structures, but minor problems in complex structures.<br>Several errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony but they rarely obscure the meaning. |
| | FAIR TO POOR | 3-4 | Uses limited range of structures<br>Major problems in simple and complex structures.<br>Frequent errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony<br>These errors frequently impede the meaning. |
| | POOR | 1-2 | Little or no mastery of sentence structures<br>Dominated by errors that blocks the meaning. |

| | | | |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of vocabulary with a sophisticated control of lexical features.<br>Effectively uses infrequent lexical items, but there may be occasional errors in word choice and collocations<br>Produces no or rare errors in spelling and word formation. |
| | GOOD TO AVERAGE | 5-6 | Uses an adequate range of vocabulary<br>Attempts to use less frequent vocabulary, but with same inaccuracy.<br>Makes some errors in spelling and word formation, but they do not obscure communication. |
| | FAIR TO POOR | 3-4 | Uses a limited range of vocabulary which may be used repetitively or inappropriately.<br>Makes frequent errors of word formation and spelling<br>Errors may obscure meaning |
| | POOR | 1-2 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but meaning is incomprehensible. |

| | |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes a totally memorized response |

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE- TASK3

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 7-8 | Fully addresses all the parts of the task and the topic,<br>Presents a well- developed position to the question with relevant main ideas and draw appropriate conclusions.<br>Thorough development of main ideas logically supported with explanations and examples. |
| | GOOD TO AVERAGE | 5-6 | Addresses all parts of the task although some parts may not be elaborated adequately.<br>Presents a position in answer to the question although the conclusions may be unclear or repetitive<br>Presents relevant main ideas, but they are generally limited, not adequately developed.<br>There may be irrelevant details. |
| | FAIR TO POOR | 3-4 | Addresses the task only partially<br>Expresses a position but it is not clear<br>Present some or few main ideas but these are mostly undeveloped, repetitive or irrelevant. |
| | POOR | 1-2 | Answer is hardly related or completely unrelated to the task, does not express a position, there is little or no development of ideas. |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 7-8 | Ideas and information are well-organized and logically sequenced with beginning, development and ending.<br>Ideas are consistent and connected. Transition between ideas is well managed with effective use of cohesive devices (linking words, referencing and substitution)<br>Effectively manages paragraphing. |
| | GOOD TO AVERAGE | 5-6 | Ideas and information are presented with some organization.<br>There is an overall unity and progression.<br>Ideas are generally consistent and connected.<br>The logical sequence and connection of ideas may be affected by limited, inaccurate or over-use of cohesive devices.<br>May not always use referencing appropriately and clearly.<br>Uses paragraphing but not always logically. |
| | FAIR TO POOR | 3-4 | Ideas and information are presented but not organized logically. There is little progression in the response.<br>Uses some basic cohesive devices but these are often inaccurate or inappropriate.<br>Ideas are often disconnected, repetitive or mechanical because of lack of referencing, substitution and linking words.<br>May not write in paragraphs or the paragraphs may be confusing. |
| | POOR | 1-2 | Weak or no organization, sequencing; and no connection and consistency between ideas. |

|  |  |  |  |
|---|---|---|---|
| GRAMMAR | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of complex structures<br>Makes few errors of subject-verb agreement, tense, number, word order, pronouns, negation , vowel harmony and case-markers. |
|  | GOOD TO AVERAGE | 5-6 | Uses effective but simple structures.<br>Attempts to use complex structures, but minor problems in complex structures.<br>Several errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony but they rarely obscure the meaning. |
|  | FAIR TO POOR | 3-4 | Uses limited range of structures<br>Major problems in simple and complex structures.<br>Frequent errors of subject-verb agreement, tense, number, word order, pronouns, negation, case markers and vowel harmony<br>These errors frequently impede the meaning. |
|  | POOR | 1-2 | Little or no mastery of sentence structures<br>Dominated by errors that blocks the meaning. |

|  |  |  |  |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of vocabulary with a sophisticated control of lexical features.<br>Effectively uses infrequent lexical items, but there may be occasional errors in word choice and collocations<br>Produces no or rare errors in spelling and word formation. |
|  | GOOD TO AVERAGE | 5-6 | Uses an adequate range of vocabulary<br>Attempts to use less frequent vocabulary, but with same inaccuracy.<br>Makes some errors in spelling and word formation, but they do not obscure communication. |
|  | FAIR TO POOR | 3-4 | Uses a limited range of vocabulary which may be used repetitively or inappropriately.<br>Makes frequent errors of word formation and spelling<br>Errors may obscure meaning |
|  | POOR | 1-2 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but meaning is incomprehensible. |

|  |  |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes a totally memorized response |

*Initial Draft (Used in the First Phase)*

WRITING RUBRIC FOR TEST OF TURKISH AS A FOREIGN LANGUAGE

| | | | |
|---|---|---|---|
| CONTENT | EXCELLENT TO VERY GOOD | 7-8 | Fully addresses all requirements of the task,<br>A clear main idea,<br>Thorough development of thesis/ideas logically supported with reasons and examples. |
| | GOOD TO AVERAGE | 5-6 | Covers the requirements of the task,<br>Satisfactory ideas and they are mostly relevant to the topic although development and support may be limited |
| | FAIR TO POOR | 3-4 | Attempts to address the task but does not cover all the key features,<br>Inadequate development of the topic, ideas may be unclear, irrelevant or repetitive.<br>The tone and register may be inappropriate. |
| | POOR | 1-2 | Answer is hardly related or completely unrelated to the task, irrelevant support or no development. |

| | | | |
|---|---|---|---|
| ORGANIZATION | EXCELLENT TO VERY GOOD | 7-8 | Ideas and information are well-organized and logically sequenced with beginning, development and ending.<br>Ideas are consistent and connected. Transition between ideas is well managed with effective use of cohesive devices. |
| | GOOD TO AVERAGE | 5-6 | Ideas are loosely organized but main ideas stand out. Beginning, development and ending may be imbalanced. Logical sequence and connection of ideas may be affected by limited use of cohesive devices. |
| | FAIR TO POOR | 3-4 | Ideas are often inconsistent and disconnected.<br>Wrong or irrelevant use of cohesive devices<br>Lacks logical sequencing |
| | POOR | 1-2 | Weak or no organization, sequencing and no connection and consistency between ideas. |

| | | | |
|---|---|---|---|
| GRAMMAR | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of complex structures<br>Makes few errors of subject-verb agreement, tense, number, word order/function, pronouns, negation , vowel harmony and case-markers. |
| | GOOD TO AVERAGE | 5-6 | Uses effective but simple structures.<br>Attempts to use complex structures, but minor problems in complex structures.<br>Several errors of subject-verb agreement, tense, number, word order/function, pronouns, negation, case markers and vowel harmony but they rarely obscure the meaning. |
| | FAIR TO POOR | 3-4 | Uses limited range of structures<br>Major problems in simple and complex structures.<br>Frequent errors of subject-verb agreement, tense, number, word order/function, pronouns, negation, case markers and vowel harmony<br>These errors frequently impede the meaning. |
| | POOR | 1-2 | Little or no mastery of sentence structures |

| | | | |
|---|---|---|---|
| | | | Dominated by errors that blocks the meaning. |

| | | | |
|---|---|---|---|
| VOCABULARY | EXCELLENT TO VERY GOOD | 7-8 | Uses a wide range of vocabulary with a sophisticated control of lexical features.<br>Effectively uses infrequent lexical items, but there may be occasional errors in word choice and collocations<br>Produces no or rare errors in spelling and word formation. |
| | GOOD TO AVERAGE | 5-6 | Uses an adequate range of vocabulary<br>Attempts to use less frequent vocabulary, but with same inaccuracy.<br>Makes some errors in spelling and word formation, but they do not obscure communication. |
| | FAIR TO POOR | 3-4 | Uses a limited range of vocabulary which may be used repetitively or inappropriately.<br>Makes frequent errors of word formation and spelling<br>Errors may obscure meaning |
| | POOR | 1-2 | Has little or no knowledge of Turkish vocabulary and word form.<br>Uses only a few relevant words, but meaning is incomprehensible. |

| | |
|---|---|
| 0 | Does not attempt the task in any way<br>Writes a totally memorized response |

APPENDIX D

TASK EVALUATION FORM

**Surname, Name** ____________________________
**Class:** ____________________________

Were the instructions clear?

| | | |
|---|---|---|
| YES<br>□ | NO<br>□ | If no, which part? |

Was the prompt clear?

| | | |
|---|---|---|
| YES<br>□ | NO<br>□ | If no, which part? |

Was the time enough to complete the task?

| | | |
|---|---|---|
| YES<br>□ | NO<br>□ | If no, any suggested duration? |

Was the task similar to the tasks you do in real life?

| | |
|---|---|
| YES<br>□ | NO<br>□ |

Was the task similar to the tasks you do in class?

| | |
|---|---|
| YES<br>□ | NO<br>□ |

Was the topic interesting to you?

| | | |
|---|---|---|
| YES<br>□ | NO<br>□ | Suggestions: |

Please evaluate the difficulty level of the task:

| | | | |
|---|---|---|---|
| Easy<br>□ | Moderate<br>□ | Difficult<br>□ | Very difficult<br>□ |

Was the minimum number of words appropriate to complete the task?

| | | |
|---|---|---|
| YES<br>□ | NO<br>□ | Suggestions: |

# APPENDIX E

## EXPERT OPINION FORM

**Part A**

For each of the items below, circle the number that reflects your opinion on a four-point scale where:
1 = Strongly disagree, 2 = Disagree, 3 = Agree 4 = Strongly agree

| | | |
|---|---|---|
| **INSTRUCTIONS** | | |
| | Instructions are clear. | 1 2 3 4 |
| | Instructions are adequate. | 1 2 3 4 |
| | Instructions are relevant. | 1 2 3 4 |
| **TEST CONTENT** | | |
| | **Task A assesses the ability to narrate a personal experience in the form of a story.** | 1 2 3 4 |
| | Task A is a good task to assess the ability to communicate ideas in written form in an academic context. | 1 2 3 4 |
| | The type of response required from the student in Task A is appropriate in an academic context. | 1 2 3 4 |
| | Task A represents the domain of academic Turkish required for studies at a Turkish-medium university. | 1 2 3 4 |
| | The time allowed for Task A is sufficient. | 1 2 3 4 |
| | **Task B assesses the ability to summarize and report factual information which is provided through graphs, charts or tables.** | 1 2 3 4 |
| | Task B is a good task to assess the ability to communicate ideas in written form in an academic context. | 1 2 3 4 |
| | The type of response required from the student in Task B is appropriate in an academic context. | 1 2 3 4 |
| | Task B represents the domain of academic Turkish required for studies at a Turkish-medium university. | 1 2 3 4 |
| | The time allowed for Task B is sufficient. | 1 2 3 4 |
| | **Task C assesses the ability to develop an argument in support of or against a point of view and back it up with reasons and examples in written form.** | 1 2 3 4 |
| | Task C is a good task to assess the ability to communicate ideas in written form in an academic context. | 1 2 3 4 |
| | The type of response required from the student in Task C is appropriate in an academic context. | 1 2 3 4 |
| | Task C represents the domain of academic Turkish required for studies at a Turkish-medium university. | 1 2 3 4 |
| | The time allowed for Task C is sufficient. | 1 2 3 4 |
| | | 1 2 3 4 |
| **OVERALL** | | |
| | The tasks are ordered from easy to more difficult | 1 2 3 4 |
| | The criteria for scoring are adequate to reflect performance at different levels. | 1 2 3 4 |
| | | 1 2 3 4 |

**Part B**

If you have any suggestions to improve the tasks, please indicate them below.

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

## APPENDIX F

## PARTICIPANT PROFILE FORM

### I. PERSONAL INFORMATION (Will Remain Confidential)

Last Name, First Name:______________________________________________

E-mail address:______________________________________________

Sex: Female______________ Male:__________

Date of Birth:__________________Place of Birth: City:________ Country:__________________

Occupation:______________________________________________

Highest Level of Schooling: Secondary__________High school____________University ____________

### II. LINGUISTIC INFORMATION

Mother Tongue: ______________________________________________

Language of Education:______________________________________________

Primary School:________________Secondary School:______________________________

High School:__________________University:______________________________

Age & place of first exposure to Turkish: ______________________________

How long have you been learning Turkish? (*e.g. for 8 months*)______________________________

How long have you been living in Turkey? ______________________________

How often do you use Turkish? (e.g., *5 hours a week*) ______________________________

What language do you generally use? Home:____________Work:________________ Social:_________

**III. SECOND/FOREIGN LANGUAGE(S):**

**Please refer to the following table of Common European Framework Reference Levels to identify your level of proficiency in the languages you know. Note the second/foreign language you know and write your level in the space provided for each language.**

**Turkish:** ____________________________________

**Second/Foreign Language 2:** ____________________________________

**Second/Foreign Language 3:** ____________________________________

| | | |
|---|---|---|
| **Proficient user** | **C2** | Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express myslef spontaneously, very fluently and precisely in complex situations. |
| | **C1** | Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express myself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects. |
| **Independent user** | **B2** | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in my field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. |
| | **B1** | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations. Can produce simple connected text on topics of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. |
| **Basic user** | **A2** | Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe my background and immediate environment in simple terms. |
| | **A1** | Can understand and use familiar everyday expressions and very basic phrases. Can introduce myself and others and can ask and answer questions about personal details. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help. |

**IV. TURKISH LANGUAGE PROFICIENCY**

Have you taken any formal instruction in Turkish? If so, where?______________________________________________

_______________________________________________

Have you taken any Turkish proficiency/placement test? If so, please indicate the name of the test and the result._______

_______________________________________________

How would you rate your linguistic ability in Turkish in the following areas? Please put a tick on the relevant box for each language skill.

| | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| **Writing** | I can write a short, simple postcard. I can fill in forms with personal details.<br>□ | I can write short, simple notes and messages. I can write personal letters.<br>□ | I can write simple connected text on topics of personal interest. I can write personal letters describing experiences and impressions.<br>□ | I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.<br>□ | I can express myself in clear, well- structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind<br>□ | I can write clear, smoothly flowing text in an appropriate style. I can write complex subjects reports or articles which present a case with an effective logical structure. I can write summaries and reviews of professional or literary works.<br>□ |

# APPENDIX G

# MANY-FACET RASCH ANALYSIS OUTPUT

## RATER MEASUREMENT REPORT

```
+----------------------------------------------------------------------------------------------------------------------------
--+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation | Exact Agree. |
|
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %  Exp % | N rater
|
|--------------------------------+--------------+---------------------+------+-------------+--------------+------------------
---|
|   910     192      4.74   4.44 |    .18   .09 | 1.14  1.3  1.09   .8 |  .85 |   .88   .89 |  37.5   37.2 | 3 Y
|
|  1008     212      4.75   4.61 |   -.03   .08 |  .92  -.8   .90  -.9 | 1.08 |   .90   .89 |  48.1   36.7 | 1 F
|
|  1067     220      4.85   4.71 |   -.15   .09 |  .82 -1.8   .86 -1.2 | 1.15 |   .91   .91 |  42.7   38.7 | 2 T
|
|--------------------------------+--------------+---------------------+------+-------------+--------------+------------------
---|
|   995.0   208.0    4.78   4.58 |    .00   .09 |  .96  -.5   .95  -.5 |      |   .90       |              | Mean (Count: 3)
|
|    64.8    11.8     .05    .11 |    .14   .00 |  .13  1.3   .10   .9 |      |   .01       |              | S.D. (Population)
|
|    79.3    14.4     .06    .14 |    .17   .00 |  .16  1.6   .13  1.1 |      |   .02       |              | S.D. (Sample)
|
+----------------------------------------------------------------------------------------------------------------------------
---+
Model, Populn: RMSE .09  Adj (True) S.D. .11  Separation 1.26  Strata 2.01  Reliability (not inter-rater) .61
Model, Sample: RMSE .09  Adj (True) S.D. .15  Separation 1.70  Strata 2.60  Reliability (not inter-rater) .74
Model, Fixed (all same) chi-square:  7.5  d.f.: 2  significance (probability): .02
Model,  Random (normal) chi-square:  1.6  d.f.: 1  significance (probability): .21
Inter-Rater agreement opportunities: 312  Exact agreements: 134 =  42.9%  Expected:  117.2 =  37.6%
-----------------------------------------------------------------------------------------------------------------------------
------------------
```

## TASK MEASUREMENT REPORT

```
+-----------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |                      |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N task               |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|  1388     312      4.45   4.15 |    .54   .07 | 1.08   .9  1.09   .9 |  .90 |   .88   .89 | 1 graph              |
|  1597     312      5.12   5.03 |   -.54   .07 |  .83 -2.1   .80 -2.2 | 1.18 |   .91   .89 | 2 essay              |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|  1492.5  312.0     4.78   4.59 |    .00   .07 |  .95  -.6   .94  -.6 |      |   .90       | Mean (Count: 2)      |
|   104.5     .0      .33    .44 |    .54   .00 |  .12  1.6   .14  1.6 |      |   .02       | S.D. (Population)    |
|   147.8     .0      .47    .62 |    .77   .00 |  .18  2.2   .20  2.3 |      |   .02       | S.D. (Sample)        |
+-----------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .07  Adj (True) S.D. .54  Separation 7.59  Strata 10.46  Reliability .98
Model, Sample: RMSE .07  Adj (True) S.D. .76  Separation 10.79  Strata 14.71  Reliability .99
Model, Fixed (all same) chi-square:  117.3  d.f.: 1  significance (probability): .00
-------------------------------------------------------------------------------------------------------------------
```

CRITERIA MEASUREMENT REPORT

```
+------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |                        |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N criteria             |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|   715     156     4.58   4.34 |    .31   .10 | 1.40  3.0  1.34  2.5 |  .57 |   .87   .90 | 1 content              |
|   743     156     4.76   4.56 |    .03   .10 |  .93  -.5   .99   .0 | 1.00 |   .90   .90 | 2 organisation         |
|   753     156     4.83   4.64 |   -.07   .10 |  .82 -1.6   .81 -1.5 | 1.20 |   .91   .90 | 3 language             |
|   774     156     4.96   4.81 |   -.28   .10 |  .67 -3.1   .64 -3.0 | 1.37 |   .92   .90 | 4 vocabulary           |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|   746.3   156.0    4.78   4.59 |    .00   .10 |  .95  -.6   .94  -.5 |      |   .90       | Mean (Count: 4)        |
|    21.2      .0     .14    .17 |    .21   .00 |  .27  2.3   .26  2.1 |      |   .02       | S.D. (Population)      |
|    24.5      .0     .16    .20 |    .24   .00 |  .32  2.7   .30  2.4 |      |   .02       | S.D. (Sample)          |
+------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .10  Adj (True) S.D. .19  Separation 1.86  Strata 2.82  Reliability .78
Model, Sample: RMSE .10  Adj (True) S.D. .22  Separation 2.23  Strata 3.31  Reliability .83
Model, Fixed (all same) chi-square:  17.9  d.f.: 3  significance (probability): .00
Model,  Random (normal) chi-square:  2.6  d.f.: 2  significance (probability): .27
--------------------------------------------------------------------------------------------------------------------
```

CATEGORY STATISTIC (GRAPH INTERPRETATION TASK)

```
+-----------------------------------------------------------------------------------------------------------------------+
|           DATA                  |   QUALITY CONTROL |RASCH-ANDRICH|  EXPECTATION  |  MOST  |  RASCH-   | Cat|Response  |
|      Category Counts       Cum.|  Avge  Exp. OUTFIT| Thresholds  |  Measure at   |PROBABLE| THURSTONE|PEAK|Category  |
|Score Total      Used    %     % |  Meas  Meas  MnSq |Measure  S.E.|Category  -0.5 |  from  |Thresholds|Prob|  Name    |
|---------------------------------+-------------------+-------------+---------------+--------+----------+----+----------|
|  1      7         7    2%    2%| -3.53  -3.95  1.4 |             |( -6.22)       |  low   |   low    |100%| Poor1    |
|  2     32        32   10%   13%| -2.98  -2.99  1.2 | -5.06    .43|  -4.11   -5.32|  -5.06 |  -5.17   | 57%| Poor2    |
|  3     78        78   25%   38%| -1.36  -1.47  1.1 | -3.11    .24|  -1.93   -3.03|  -3.11 |  -3.07   | 61%| weak     |
|  4     67        67   21%   59%|  -.46   -.37  1.2 |  -.74    .18|   -.05    -.89|   -.74 |   -.82   | 50%| fair     |
|  5     33        33   11%   70%|   .68    .94   .9 |   .94    .22|   1.17     .61|    .94 |    .72   | 36%| average  |
|  6     41        41   13%   83%|  2.19   2.28  1.2 |  1.46    .24|   2.24    1.70|   1.46 |   1.63   | 46%| good     |
|  7     28        28    9%   92%|  3.16   2.94   .8 |  3.01    .22|   3.40    2.79|   3.01 |   2.84   | 37%| very good|
|  8     26        26    8%  100%|  4.14   4.13   .9 |  3.50    .28|(  4.90)   4.20|   3.50 |   3.87   |100%| excellent|
+-------------------------------------------------------------------(Mean)---------(Modal)--(Median)------------------+
```

CATEGORY STATISTICS (ESSAY TASK)

```
+------------------------------------------------------------------------------------------------------------------------+
|          DATA                   |   QUALITY CONTROL |RASCH-ANDRICH|  EXPECTATION   |  MOST   |  RASCH-   | Cat|Response  |
|     Category Counts         Cum.|  Avge  Exp. OUTFIT| Thresholds  |  Measure at    |PROBABLE| THURSTONE|PEAK|Category  |
|Score Total      Used     %     % |  Meas  Meas  MnSq |Measure  S.E.|Category  -0.5  |  from   |Thresholds|Prob|  Name    |
|---------------------------------+-------------------+-------------+----------------+---------+----------+----+----------|
|  1       1         1    0%    0%| -2.77  -2.98  1.0 |             |( -6.94)        |   low   |   low    |100%| Poor1    |
|  2      22        22    7%    7%| -2.72  -2.44   .6 | -5.85   1.01|  -4.23   -5.92|  -5.85  |  -5.88   | 71%| Poor2    |
|  3      47        47   15%   22%| -1.17  -1.16   .9 | -2.61    .28|  -1.77   -2.81|  -2.61  |  -2.71   | 54%| weak     |
|  4      64        64   21%   43%|   .09    .04   .8 |  -.82    .20|   -.11    -.89|   -.82  |   -.86   | 49%| fair     |
|  5      51        51   16%   59%|   .95    .93   .9 |   .72    .18|   1.22     .59|    .72  |    .63   | 43%| average  |
|  6      38        38   12%   71%|  2.09   2.23   .8 |  1.82    .22|   2.35    1.79|   1.82  |   1.80   | 41%| good     |
|  7      40        40   13%   84%|  3.75   3.56   .6 |  2.90    .23|   3.54    2.91|   2.90  |   2.89   | 42%| very good|
|  8      49        49   16%  100%|  4.62   4.61  1.0 |  3.83    .22|(  5.14)   4.40|   3.83  |   4.11   |100%| excellent|
+-------------------------------------------------------------------(Mean)---------(Modal)--(Median)-----------------+
```

# REFERENCES

Ahmadi, A., & Mansoordehghan, S. (2015). Task type and prompt effect on test performance: A focus on IELTS academic writing tasks. *Journal of Teaching Language Skills*, *6*(3), 1-20.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analysis for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293.

Becker, A. (2010). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal*, *22*(1), 113-130.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

Brindley, G. (2000). Task difficulty and task generalisability in competency-based writing assessment. *Studies in Immigrant English Language Assessment*, *1*, 125-157.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Council of Europe (2001). *Common European reference framework for languages*. Strasbourg, France: Author. Retrieved from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents_intro/common_framework.html

Council of Europe (2008). *The CEFR Grid for Writing Tasks*. Strasbourg, France: Author. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/CEFRWritingGridv3_1_analysis.doc

Council of Europe (2009). *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)*. Strasbourg, France: Author. Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1-8.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 writing framework: A working paper (TOEFL Monograph Series, Report no. 18). Princeton, NJ: Educational Testing Service.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teachers guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*(3), 140-150.

Davies A, Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing*. Studies in Language Testing, Vol. 7. Cambridge: UCLES/ Cambridge University Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88-115.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270-292.

Ellis, M., Chong, S., & Choy, Z. (2013). IELTS as an indicator of written proficiency levels: A study of student teachers at the National Institute of Education, Singapore. *International Journal of Educational Research*, *60*, 11-18.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93-112.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England & New York, NY: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5-29.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507-531.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7,* 47–84.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective.* Harlow: Pearson Education.

Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish.* (Unpublished master’s thesis). Boğaziçi University, Istanbul, Turkey.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49-68.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1-33.

Hayes, R. J. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Erlbaum.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64-86.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237-263.

Hyland, K. (2002). *Teaching and Researching Writing*. London: Longman.

Ishikawa, T. (2006). The effect of task complexity and language proficiency on task-based language performance. *The Journal of Asia TEFL*, *3*(4), 193-225.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Khalifa, H. (2009). Aligning Cambridge ESOL Examinations to the CEFR: Issues & Practice. *Cambridge ESOL: Research Notes*, *37*, 10-14.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing 26, Cambridge: Cambridge University Press.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 5, 1–36.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale.* Frankfurt, Germany: Peter Lang.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing,17*(1), 48-60.

Kuiken, F., Vedder, I., & Matters, M. (2007). Cognitive task complexity and linguistic performance in French L2 writing. *Investigating Tasks in Formal Language Learning*, *20*, 117.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418.

Lee, S., & Kim, N. (2007). Syntactic and lexical variation in rhetorical tasks of low-intermediate EFL college students. *English Teaching*, *62*(3), 357-373.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131-166.

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series*, *2005*(1), i-76.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness, *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2014). FACETS (Version 3.71.4) [Computer software]. Chicago, IL: MESA Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series*, *1996*(1), i-18.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, *4*(1), 43-66.

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, *15*(2), 187-215.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

North, B. (2004). Relating assessments, examinations, and courses to the CEF. *Insights from the Common European Framework*, 77-90.

O'Sullivan, B. (2005). *A practical introduction to using FACETS in language testing research*. Unpublished manuscript, University of Roehampton, London, UK.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218-233.

Osburne, A. G., & Mulling, S. (1994). Essay prompts and the ESOL student. *Issues in Applied Linguistics*, *5*(1).

Pallant, J. F. (2007). SPSS Survival Manual: A step-by-step guide to data analysis with SPSS. New York: McGraw-Hill.

Paltridge, B. (2004). Academic writing. *Language Teaching*, *37*(02), 87-105.

Papp, S., & Salamoura, A. (2009). An exploratory study into linking young learners' examinations to the CEFR. *Research Notes*, *37*, 15-22.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Papers in TESOL & Applied Linguistics*, *4*, 1-21.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*(2), 111-129.

Plakans, L. M. (2007). *Second language writing and reading-to-write assessment tasks: A process study.* (Unpublished doctoral dissertation). The University of Iowa, Iowa City, IA.

Reed, W. M. (1992). The effects of computer-based writing tasks and mode of discourse on the performance and attitudes of writers of varying abilities. *Computers in Human Behavior*, *8*(1), 97-119.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Sakyi, A. (2001). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 130–153). Cambridge: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge: Cambridge University Press.

Spack, R. (1988). Initiating ESL students into the academic discourse community: How far should we go? *TESOL Quarterly*, *22*(1), 29-51.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261.

Taylor, L. (2011). *Examining Speaking: Research and practice in assessing second language speaking* (Vol. 30). Cambridge: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3-12.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *Modern Language Journal*, 171-184.

Webb, N.M., & Shavelson, R.J. (2005). Generalizability theory: Overview. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 717-719). Chichester: John Wiley & Sons Ltd.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, *37*(2), 345-354.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing*, *24*(2), 251-286.

Yang, H. C. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes*, *31*(3), 174-187.

Yu, G., Rea-Dickins, P., & Kiely, R. (2012). The cognitive processes of taking IELTS academic writing task 1. *IELTS Research Reports Volume 11, 2012, 2nd edition*, 1.

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325.