DETERMINING LINGUISTIC DEMANDS OF TEXTS FOR EAP PROFICIENCY TESTS USING AUTOMATED ANALYSES AND EXPERT JUDGMENT

BEKİR ATEŞ

BOĞAZİÇİ UNIVERSITY

2018

DETERMINING LINGUISTIC DEMANDS OF TEXTS FOR EAP PROFICIENCY TESTS USING AUTOMATED ANALYSES AND EXPERT JUDGMENT

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

Bekir Ateş

Boğaziçi University

2018

DECLARATION OF ORIGINALITY

I, Bekir Ateş, certify that

- I am the sole author of this 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: 11.04.2018

ABSTRACT

Determining Linguistic Demands of Texts for EAP Proficiency Tests

Using Automated Analyses and Expert Judgment

The purpose of this study is to investigate the use of expert judgments and automated textual analysis tools as instruments in generating context validity evidence for the use of reading test texts for EAP proficiency exams. Based on the linguistic task demands outlined in Khalifa and Weir's (2009) validation framework for reading tests, the study aimed to explore the text features that influenced experts' judgments on the difficulty and suitability of texts for an EAP reading test. Results obtained from the analysis of 10 texts through 24 automated textual analysis indices were correlated with the judgments of experts on different textual features of the texts in order to determine the automated indices that could readily replace expert judgments. Textual analyses results for 120 texts from four corpuses (a corpus of IELTS reading test texts and course books of three universities, namely İstanbul Şehir University, Boğaziçi University, and University of Bedfordshire) were compared to find the similarities and differences between the corpuses. Finally, through a descriptive analysis of the three university corpuses (90 texts of about 800 words each) optimal ranges of textual analysis index scores representing the majority of the university texts were offered. The findings from this study provide guidance to test developers in their efforts to generate context validity evidence by means of expert judgments and automated textual analysis tools.

ÖZET

Akademik Amaçlı İngilizce Yeterlilik Sınavlarının Dilsel Gereksinimlerinin Uzman Görüşü ve Otomatik Metin Analizleriyle Belirlenmesi

Bu araştırma akademik amaçlı İngilizce yeterlilik sınavlarında kullanılacak okuma metinlerinin bağlam geçerliliği için kanıt toplanmasında uzman görüşlerinin ve otomatik metin analiz araçlarının kullanımını incelemektir. Khalifa ve Weir'in (2009) geçerlilik teorisinde belirtilen metin gereksinimleri göz önünde bulundurulmuş ve buna bağlı olarak uzmanların bir metnin akademik bir İngilizce yeterlilik sınavının okuma bölümünde kullanılması konusunda yaptıkları metin zorluğu ve uygunluğuna dair yorumlarını etkileyen metinsel özellikler bulunmaya çalışılmıştır. 24 otomatik analiz endeksiyle 10 metnin incelenmesinden elde edilen skorların uzman yorumlarıyla ilişkisi incelenerek uzman yorumlarının yerini tutabilecek otomatik analizler belirlenmeye çalışılmıştır. IELTS okuma sınavları ve üç üniversitenin (Boğaziçi Üniversitesi, İstanbul Şehir Üniversitesi ve Bedfordshire Üniversitesi) ders kitaplarından alınan toplamda 120 metinlik dört korpus otomatik analiz sonuçları bakımından benzerlikleri ve farklılıklarının bulunması amacıyla karşılaştırılmıştır. Son olarak, her biri yaklaşık 800 kelimeden oluşan 90 metinlik üç üniversite korpusu analiz edilerek üç üniversiteden gelen metinlerin çoğunluğunu temsil eden metinsel analiz endeks skor aralıkları sunulmuştur. Bu çalışmanın bulguları uzman görüşü ve otomatik metin analiz araçları kullanılarak bağlam geçerliliği için kanıt toplanması konusunda test yazarlarına yol göstermektedir.

ACKNOWLEDGEMENTS

I feel indebted to many people for their support in the completion of this thesis. First of all, I would like to express my deepest gratitude to my advisor Assist. Prof. Aylin Ünaldı, who was much more than an advisor to me. There were times when I was ready to give up, and she was always there to put me back on track with her patience, encouragement, creativity, and friendliness. I do not think there are many thesis advisors out there who would bother to spend an hour talking about the health situation of an advisee's son. Aylin Hoca will remain a source of inspiration to me for the rest of my life.

I would also like to thank my dear wife, Nagehan Akgün Ateş, for her endless love and support in times when I felt frustrated by my lack of progress. I could not have finished this thesis without her loving kisses she gave me every time she brought a mug of coffee to my desk.

I owe an apology to my dear son, Barış Efe, for the times I neglected him working on this thesis. I promise him that I will do my best to make up for all that time.

I should also thank my colleague Mehmet Akıncı, who was never short of motivational speeches to lift up my mood. I am also grateful to him for providing me with guidance in statistical analyses.

Finally, I would like to thank Mehtap İnce, İlke Büyükduman, and all the other colleagues who contributed to the completion of this study by not only providing me with valuable insight but also agreeing to participate in the study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDIX TABLES

CHAPTER 1

INTRODUCTION

1.1 Background to the study

Assessment can be defined as the process of collecting information about the abilities, knowledge or readiness of the people assessed and then making decisions based on the information collected. In the case of high stakes testing, score-based interpretations and decisions might have a profound influence on people's lives. Bachman and Palmer (1996) state that test writers need to be able to justify their score based interpretations by generating evidence (p. 20). This process of collecting evidence for the defensibility of the use of a test and the justifiability of the interpretations made base on is called test validation. The justifiability of the score based inferences from a test depends on the amount of validity evidence that test designers can present (Messick, 1996).

The assessment of reading comprehension is crucially significant in a broad range of educational contexts, and there is an obvious need for expertise in this area. Alderson (2000) points out that reading comprehension is the result of the interplay between reader variables (e.g. intelligence, background knowledge, reader strategies etc.) and text variables (e.g. length, lexical complexity, discourse mode, abstractness etc.). In order to make valid interpretations regarding the reading ability of test takers, developers of reading tests need to take into account these factors. Alderson (2000) warns test developers that, in text selection process, they have to be aware of textual factors that influence comprehension.

In the case of EAP reading tests, one type of validity evidence, as stated by Khalifa and Weir (2009), concerns context validity. According to McNamara (2000), context validity is 'the extent to which the test appropriately samples from the domain of knowledge or skills relevant to performance in the criterion.' Bachman and Palmer (1996) suggest that the tasks used in an exam must adequately represent the target language use (TLU) domain. Choice of inappropriate texts might result in what Messick (1995) calls 'construct irrelevant variance', which he regards as a serious threat to the validity of judgments based on test results. Given the importance of validity evidence, a relevant question to ask is how to collect it.

Validation frameworks are recommended and commonly used in validation studies because they guide researchers on collecting relevant data in a principled way. Bannur, Abidin and Jamil (2015), for instance, state that using validation frameworks ensure tests with systematic decisions and more consistent results. Kahlifa and Weir's (2009) Socio-cognitive Framework for Validating Reading Tests (SFR), which will be elaborated on in the literature review section, is commonly used in studies of reading test design and validation (e.g. Green, 2014; Ilc & Stopar, 2014; Taylor, 2014). One category of context validity evidence in this framework is about the linguistic demands of the reading test tasks. Khalifa and Weir (2009) say that it is the test developers' duty to ensure that their reading test tasks are contextually relevant in terms of linguistic task demands such as the grammatical and lexical features of the task, the content knowledge required, the nature of information, the discourse mode, and the relationship between the reader and the writer. Determining the characteristics of the linguistic demands of the reading texts in the TLU domain, test developers can define specific task descriptors (i.e. test specifications). Taylor (2014) states that identifying the explicit test specifications of the relevant features

can help test developers because such specifications can act as practical quality control tools helping the production of tests that are consistent and standard across versions.

According to Weir (2005), expert opinions and document analysis can be used in the collection of information on the relevant features of the TLU domain. Expert judgments of test developers are commonly used in determining the appropriateness of texts for inclusion in reading tests. However, subjectivity involved in these judgments might be a source of undesired variation across versions of a reading test, which is a threat to the accuracy of score interpretations (Messick, 1995) and fairness of the test (Kunnan, 2014). It is also claimed that judgments of experts might sometimes end up being wrong (Biber, Conrad, Reppen, Byrd, & Helt, 2002). For this reason, expert opinions alone may not suffice as context validity evidence, especially in the case of high stakes testing. Apparently, there is a need for more research focusing on the nature of human judgments. If we can identify what makes a text difficult suitable for a specific purpose in the eyes of human raters, we can make better use of human judgments.

As more objective sources, automated textual analysis tools are also commonly used in document analysis. However, there are too many tools to choose from and the results generated through such tools may sometimes contradict with each other. Besides, as Green, Ünaldı and Weir (2010) state, there are some features (e.g. content knowledge required) that automated tools fail to capture. Consequently, relying only on automated tools in developing test specifications might lead to undesirable results as these tools may not assess every text feature. However, relying only on the judgments of the test developers may also cause problems, as tests may not be standard across versions due to the subjectivity involved. Plus, it clearly is not

practical for humans to do a task if that task can easily be done by automated textual analysis tools.

It seems that in order to utilize expert judgments and automated textual analysis tools effectively; one should analyze the nature of expert judgments and determine the automated tools that can be reliably and practically applicable in support of expert judgments. Identifying which tools can replace or support human judgment and under what conditions they work properly can provide test developers invaluable information in addition to saving a great deal of time and effort. There are a lot of studies conducted using textual analysis tools. Some of these studies have been conducted to compare the effectiveness of tools in assigning texts to correct grade levels or checking the differences of texts in different grade levels (e.g. Crossley & McNamara, 2008; Dufty, Graesser, Louwerse, & McNamara, 2006; Plakans & Bilki, 2016). There are also studies that the developers of automated tools conducted in order to validate the use of their tools by comparing them against human judgment (e.g. Kyle & Crossley, 2015). However, there seems to be a lack of research that compares a variety of tools against the judgments of experts on multiple features of texts. It seems that there is a need for studies that use a variety of tools to analyze different textual features and compare the findings with the judgments of experts on the same features. This way it would be possible to identify the tools that agree with the judgments of the experts on certain textual features and thus could replace expert judgments. Furthermore, there seems to be a lack of research for the use of automated analysis tools in the development of tests of academic English at university contexts. A comparison of documents from university course books using textual analysis tools may shed light into the similarities shared by these contexts and

thus might be used to generate some general guidelines representing a larger context of universities.

1.2 Research questions

This study aims to propose an effective way of utilizing human judgments and textual analysis tools in the generation of context validity evidence for university level EAP reading tests.

In an attempt to explore the nature of human judgments on text difficulty and suitability by analyzing the text features that seem to drive those judgments, the first question was formulated as:

Research Question 1a: What are the factors that explain the experts' judgments on the difficulty of a text?

Research Question 1b: What are the factors that explain the experts' judgments on the suitability of a text?

In order to identify the textual analysis tools that could be used in support of human judgments or even replace human judgments the second question was formulated as:

Research Question 2: Which automated tools can explain experts' judgments of a text?

With the goal of comparing the textual features of the university course book texts and an EAP reading test, the third question was formulated as:

Research Question 3: Do texts from different corpuses differ from each other with respect to the features analyzed by automated tools?

Finally, in order offer some general textual characteristics that represent the general characteristics of the corpuses from the three universities, the fourth question was formulated as:

Research Question 4: What are the optimal ranges of text characteristics that will account for the texts that are gathered from three different English medium instruction (EMI) university contexts?

1.3 Overview of the thesis

Following this chapter, the thesis will present in Chapter 2 a review of the literature on the issues of validity in general, context validation based on Khalifa and Weir's (2009) SFR, and expert judgment and automated textual analyses tools as ways of collecting context validation evidence.

In Chapter 3, the context and the participants of the study, the instruments used to collect data and the analyses conducted are explained.

Chapter 4 will present the results for each research question and the discussions of the findings.

Finally, in Chapter 5, a summary of the thesis as well as the implications of the findings and the limitations of the study are given.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will start with a discussion on the concept of validity in the field of testing. Next, Khalifa and Weir's (2009) framework for validating reading tests (SFR) and context validity in the framework will be elaborated on with emphasis on the linguistic task demands category. Finally, the issue of textual complexity, and the use of expert judgment and automated textual analysis tools to measure textual complexity for the collection of context validity evidence will be addressed.

2.2 The concept of validity in the field of language testing

It is an established fact that test designers prepare tests to collect information from the observed test performances of test takers. Using this information, they make inferences regarding the knowledge or abilities of test takers. Because such inferences influence the lives of test takers in different ways, it is necessary to develop and elaborate on theories of validity and validation in all kinds of educational testing (D'Este, 2012). Roever and McNamara point out that the validity concern is of particular importance for language assessments because language assessments are used to make important decisions about people's lives involving their access to educational opportunity, citizenship, and educational achievement (as cited in Weir, 2005). The issue of validity in language assessments has been the focus of much research and debate since the1950s.

The early conceptualizations of validity were inspired from the field of psychometrics (Walt & Steyn, 2008). Goodwin and Leech (2003) describe the common view of validity in the 1950s as follows: “A test was considered to be either valid or not as evidenced by the correlations between the test and some other ‘external’ criterion measure”. This suggests that the early views of validity regarded it as a dichotomous concept, and it was considered a static property of a test.

The early conceptualizations also involved distinct types of validities. As influential theorists of the time, for instance, Cronbach and Meehl (1955) identified four types of validity, namely predictive validity, concurrent validity, content validity and construct validity. They viewed the first two of these categories as criterion oriented procedures because these procedures involve relating the test scores to some other external criterion. Content validity concerns the question of whether the test tasks are a sample of the target situation that the test writer is interested in. Finally, construct validity is involved when the tester attempts to relate an attribute (e.g. an ability) to measurable elements. This formulation of the concept has been quite influential in the following years. For example, 1966 Standards for Educational and Psychological Testing (as cited in D’Este, 2012) included the tripartite categorization of validity: content validity, construct validity, and criterion validity (which included predictive and concurrent validities).

2.2.1 Messick’s unified view of validity

Primarily driven by the works of Messick (1980 and 1989), a unified view of validity emerged as a reaction to separate validity types in the 1980s, and it has gained common acceptance since then. Currently, it is the most commonly used and applied

conceptualization of validity (Walt & Steyn, 2008). Messick (1989) defines 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 on the basis of test scores and other modes of assessment.' This view of validity has been regarded as a major change in the definition of the concept as it has challenged the traditional definitions in several ways (Goodwin & Leech, 2003; Weir, 2005). Firstly, as the expression 'the degree to which' entails, validity does not involve a dichotomy, but rather a continuum. Messick (1995) contends that the place of an examination on the validity continuum depends on the amount of evidence collected on different validity aspects. In other words, collection of evidence supporting the adequacy of score interpretations is a continuous process, and it is required if one is to make a validity claim.

Secondly, Messick's definition implies that validity is not a quality of the test itself; instead, it is a quality of the interpretations made on the basis of test scores. As Walt and Steyn (2008) put, "Messick shifted perspectives on validity from a property of test to that of score interpretation." According to this, it would be wrong to speak of a test as being inherently valid or invalid as one can only validate the adequacy and appropriateness of the score interpretations made from the test.

Thirdly, Messick's conceptualization introduces the consequences of the use of a test as an aspect relevant to validity. Messick (1995) criticizes the traditional conceptualizations of validity for neglecting the social consequences of test interpretation and use, and suggests that a validation study should involve the consideration of actual and potential consequences of test use. Weir (2005) supports this by stating that one can better evaluate the soundness of the interpretations made from test scores about the test takers' abilities by means of gathering information on

the events after the use of a test,. Put simply, the consequences of the administration of a test can serve as evidence for the validation of score interpretations of made from that test.

Finally, and possibly most importantly, this new formulation defines validity as a unitary concept. Unlike previous conceptualizations, Messickean view of validity proposes one superordinate category (i.e. construct validity), and under this umbrella term are the aspects of construct validity, some of which were previously regarded as separate types of validities. Messick (1986) reacts to Cronbach and Meehl's (1955) long held 'trinitarian' view of validity claiming that this compartmentalized view is incomplete and creates overreliance on one type of validity due to the misconception that these types can serve as substitutable forms of evidence. Similarly, Messick (1980) criticizes the view of validity types for leading to the assumption that presenting one kind of validity evidence is enough. More recently, it is emphasized that taking validity as a multifaceted concept is a more realistic way of looking at it than promoting separate validity types that are alternatives of each other (Goodwin & Leech, 2003; Weir, 2005).

However, Messick (1995) also cautions that "to speak of validity as a unified concept does not mean that it cannot be usefully differentiated into distinct aspects." He offers six aspects of construct validity and states that these are not alternatives of each other; instead, they are complementary aspects of validity. Therefore, lack of evidence for any one of them may hamper the justifiability of the judgments based on test scores (Weir, 2005).

This shift from discrete validity types to a unitary view of validity with subcategories or aspects has been regarded as a revolutionary change (Geisinger, 1992; Walt & Steyn, 2008). The current unified view of validity has been used in a

number of works on language testing and validation frameworks (e.g. Bachman & Palmer 1996; Weir, 2005). This change is also reflected in 1999 Standards (as cited in Goodwin & Leech, 2003) where, unlike its previous versions, aspects of validity rather than discrete types of validities are emphasized.

2.2.2 Sources of invalidity

According to Messick (1989), tests either include something that should not be in thetest or leave out something that should be in the test. As such, they are usually flawed measurements. In his 1995 paper Messick identifies these imperfections as construct underrepresentation and construct irrelevant variance, and views them as two major threats to validity. Construct underrepresentation occurs when a test does not cover important dimensions or aspects of the construct; as a result, it does not fully represent the construct. This implies that the test developer should define the boundaries of the construct domain and extract a representative sample from it without leaving out any significant part. According to Weir (2005), if an important part of the construct is not represented in the test, teachers may prefer not to teach it, so that test leads to negative washback. Construct irrelevance, on the other hand, occurs when some of the variance in test scores is a result of other factors than the construct itself (e.g. guessing factor, test format etc.). Two forms of construct irrelevance are construct irrelevant difficulty, which occurs when factors other than the construct itself make the test irrelevantly difficult, and construct irrelevant easiness, which occurs when factors other than the construct itself help test takers find correct answers and get undeservedly high scores. Messick (1995) points out that the biggest part of the variance in test scores must be construct-relevant because irrelevantly low or high scores reduce the truthfulness of the interpretations made

based on those scores. For this reason, Messick (1995) states that identifying construct relevant sources of task difficulty is a necessity and it can be done through exploring cognitive processes relevant to the construct and the nature of target domain processes. It is possible to see both sources of invalidity in educational assessments (Messick, 1996), that is why the extent to which the test adequately represents the construct seems to be a primary concern for test developers to eliminate possible sources of construct irrelevance. In short, evidence on relevance and representativeness of a test is essential for the validation of the score based interpretations of any test.

2.2.3 Criticisms of Messick's model

Despite receiving common acceptance, Messickean view of validity is also criticized by some researchers on various grounds. Borsboom and Mellenberg (2004), for instance, find Messick's (1989) unitary view of validity too broad and claim that 'treating every test-related issue as relevant to the concept of validity and aiming to integrate all these issues under a single header' makes it very hard for test developers to collect evidence. They argue for a simpler definition of validity because a broad validity theory like Messick's does not provide any practically applicable direction to testers and researchers. Likewise, Knoch and Elder (2013) say that Messick's model of validity, despite being widely recognized, does not provide much practical guidance on how to proceed with the validation of a test. Walt and Steyn (2008) recommend Weir's (2005) socio-cognitive framework as a practical alternative saying that Messick's framework is difficult to operationalize because its categories are not clearly separable.

Although Messick's unitary concept of validity has been criticized for being too broad or not practically conducive to research, its basic premises are still widely accepted in the field of language testing. Today, there is consensual agreement among many researchers that validity is not a property of the test itself, but rather it is about score interpretations. There is also agreement on validity being a matter of degree rather than a dichotomous concept. Although there is controversy on what aspects to be put under one superordinate category of validity, most researchers acknowledge that validity is a single unitary concept with multiple aspects.

To recapitulate what has been said so far, the definition of validity as a concept has undergone a metamorphosis, but it has remained one of the primary concerns in the field of language assessment. Currently, it seems evident that collection of evidence on different aspects of validity is essential in order to make truthful interpretations based on test results. However, as noted before, Messick's framework is not easily applicable for research purposes. Fortunately, driven by Messickean view of validity, a number of influential validation frameworks have been developed in recent years (e.g. Bachman & Palmer, 1996; Weir, 2005).

2.3 The need for a sound and applicable validation framework

The extent to which test developers can justify their interpretations is a central concern in testing because if they fail to do that then there is no reason to use exams to make decisions about individuals (Bachman & Palmer, 1996, p.95). This implies that it is the responsibility of the test developers to generate evidence on the use of tests and the accuracy of interpretations on the basis of scores. The validity of the use of a test does not depend on what the test designers claim alone; instead, they are

required to present evidence to support their score based interpretations (Weir, 2005). This is of particular importance in the case of high stakes testing because, as Bachman and Palmer (1996, p.97) state, the decisions made from high stakes assessments are likely to affect a big number of people, and errors in such assessments are often irreversible. However, Messick (1992) claims that although many test writers accept that they are responsible for providing validity evidence for the use of their tests, not many of them actually do this (as cited in Weir, 2005). Similarly, Bachman and Palmer (1996, p.21) assert that test developers need to collect evidence to support their score based interpretations instead of simply claiming that theirs tests are valid.

Validation frameworks are recommended and commonly used in validation studies because they guide researchers on collecting relevant data in a structured way. Green (2014), for instance, states that there is increasing awareness of the value of a framework that not only guides the design of a test but also can help researchers and testers collect evidence on how testing constructs are operationalized and interpreted in practice. Similarly, Bannur et al. (2014) maintain that using validation frameworks ensures tests with systematic decisions and more consistent results. Furthermore, preparing test specifications based on a widely accepted framework can help increase standardization of tests across versions (Taylor, 2014). In short, it is evident that test developers are responsible for collecting evidence on various validity aspects in order to validate their judgments based on test scores, and doing that based on a widely accepted framework is practically useful because such frameworks guide test developers in writing clear and detailed test specifications that can insure consistency of exam across different administrations and versions.

For the purposes of the current study, the researcher uses Khalifa and Weir's (2009) Socio-cognitive Framework for Validating Reading Tests (SFR), which is a refined version of Weir's (2005) Socio-cognitive Framework for Validating Tests (SF). The following part will begin with a discussion of Weir's (2005) SF and then Khalifa and Weir's (2009) SFR will be elaborated on.

2.4 Weir's (2005) Socio-cognitive Framework for Validating Tests (SF)

Weir, in his (2005) book *Language Testing and Validation: An Evidence-Based Approach*, sets out to outline a blueprint of the types of evidence test writers must provide in order to increase defensibility of their score based interpretations. Drawing on Messick's (1989) unitary view of validity, Weir (2005) suggests a Socio-cognitive Framework for Validating Tests (SF).Weir (2005, p.15) describes validation as 'a form of evaluation where a variety of quantitative and qualitative methodologies are used to generate evidence to support inferences from test scores.' SF is widely regarded as a practical framework for collection of validity evidence on language tests because it directs researchers in terms of what kind of evidence to collect for different aspects of validity (Bannur et al., 2014; Ilc & Stopar, 2014). Green (2014), for instance, says Weir's framework is not only based on a sound theoretical model, but it also presents a practical way for generation of evidence. O'Sullivan and Weir regards SF as the first systematic approach combining social, cognitive and scoring aspects of language use with test development and validation (as cited in Taylor, 2014).

Many validation studies of reading tests have employed Weir's SF (e.g. Bannur et al., 2014; Ilc & Stopar, 2014; Taylor, 2014). Applying SF to an existing

Test of English for Academic Purposes (TEAP) with the aim of obtaining detailed and explicit descriptions of test tasks, Taylor (2014) says that such descriptions or specifications derived on the basis of SF can serve as an effective quality control tool and ensure consistency between test tasks.

It should be noted that although Weir's framework draws on Messick's unified view of validity, there is one notable difference between the two with regards to the terminology used. Unlike Messick, who uses the term construct validity to refer to the superordinate category, Weir (2005) uses the term validity as the superordinate category. Weir (2005) explains the reason for this stating that:

> It (construct validity) is often used as a superordinate term for all the validities and also to refer more specifically to the theoretical construct, in the past often expressed in terms of individual cognitive ability, on which the test is based. We prefer instead to reinstate the term validity as the superordinate category of description. (p.14)

Using the term validity, instead of construct validity, as the superordinate category, Weir tries to avoid potential confusion that might arise from the multiple meanings of construct validity in literature. Weir (2005) suggests four versions of the framework to be used in collecting validity evidence for tests of four skills (i.e. reading, listening, speaking and writing). Each version has the same main categories, except for minor differences specific to the skill it is designed for. Recently, Khalifa and Weir (2009) developed a version of SF specifically for tests of second language reading. For the purposes of this study, the next part will continue with a discussion of Khalifa and Weir's (2009) Socio-cognitive Framework for Validating Reading Tests (SFR) adapted from Weir's (2005) SF.

2.5 Khalifa and Weir's (2009) Socio-cognitive Framework for Reading (SFR)

Khalifa and Weir (2009) put test taker characteristics on the top of their framework because every other category in the framework is affected by test takers' abilities, knowledge and motivation. In addition to test taker characteristics, SFR involves five key subcategories of validity: context validity, cognitive validity, scoring validity, consequential validity, and criterion-related validity (see Figure 1).

The first category in the framework is context validity, which is also defined as content validity in the literature. Weir believes that context validity is a better name for this aspect as "the term context better accounts for the social dimension of language use" (Weir, 2005, p. 19). Context validity involves the interlocutor and linguistic demands of the tasks as well as the administrative setting in which the tasks are given. Evidence for context validity comes from three sub categories (i.e. task setting, administrative setting and linguistic demands of the task). Context validity will be discussed in more detail in the next part.

The second category in the framework is cognitive validity, which is labeled by Weir (2005) as theory based validity. Khalifa and Weir (2009, p.34) define cognitive validity as "the extent to which tasks we employ elicit the cognitive processing involved in target reading contexts beyond the test itself." The cognitive processes that are designated in Khalifa and Weir's (2009) model of reading involve different types of reading (e.g. skimming, scanning, careful reading etc.) and micro and macro level cognitive reading comprehension related processes.

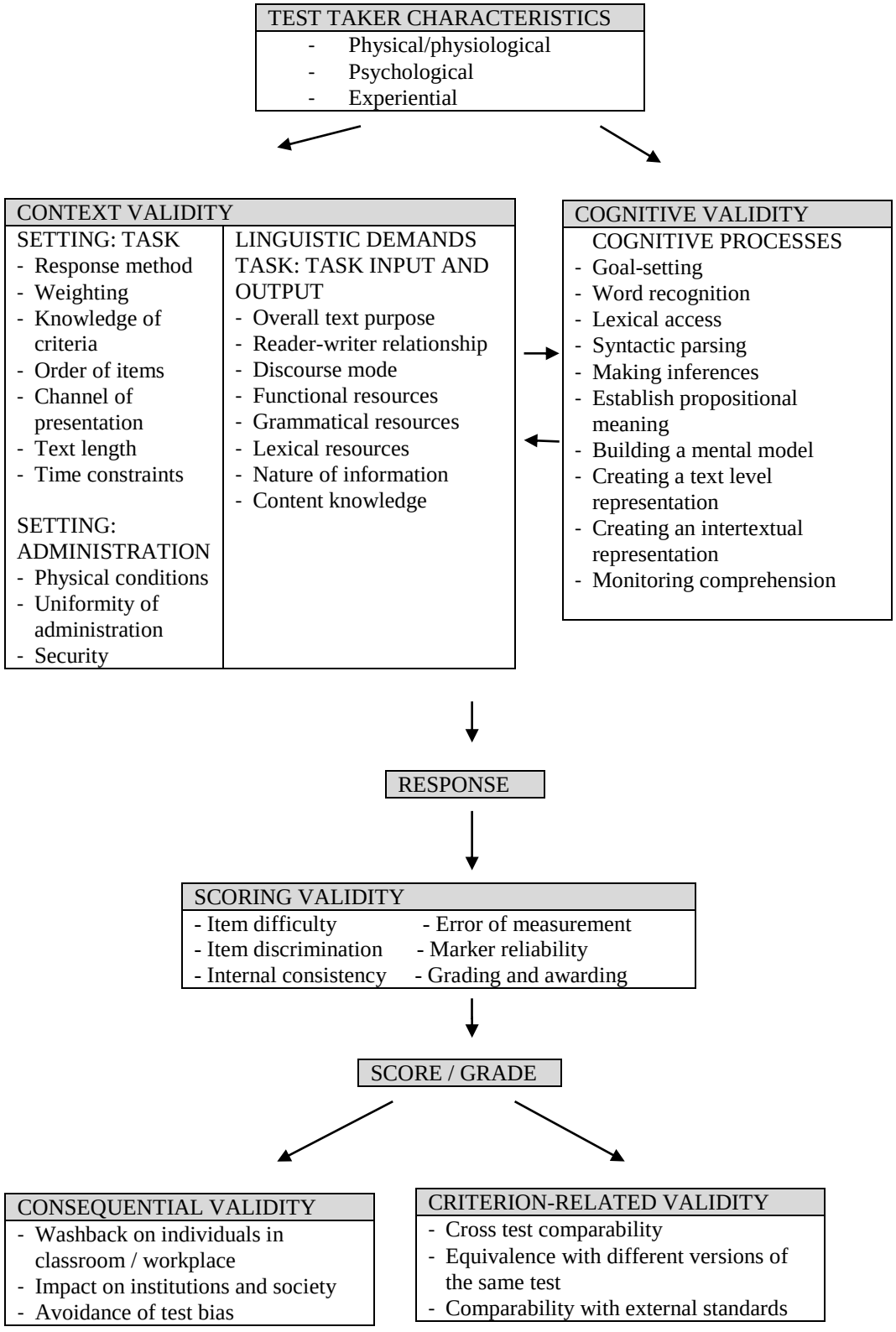

Figure 1 Socio-cognitive framework for validating reading tests (Khalifa & Weir, 2009, p.5)

The third category is scoring validity which is about collecting evidence on the reliability of the grading process. Khalifa and Weir (2009, p. 143) emphasize the important role this aspect plays saying "scoring validity is concerned with all aspects of the testing process that can impact on the reliability of test scores". Since validity is a property of score based interpretations, scoring validity is decisive in the validation process.

The fourth category is consequential validity, which has to do with the way the results of an exam impact the stakeholders and the institutions. According to Weir (2005) exams are likely to produce changes in teaching that precedes it (washback). They also have a big impact on the end users of tests (as in the use of high stakes exams for admission), and there may be a risk of bias in favor of or against some populations or groups. The investigation of these aspects can provide evidence on consequential validity.

The final category of the framework is criterion-related validity, which involves aligning exams to different proficiency levels and correlating exams scores with external criteria.

Weir (2005) emphasizes that these subcategories of validity are complementary in nature and that no single subcategory in the framework is superior to the others. Of these five subcategories, the first two are regarded as components of a priori test validation, while the following three are components of a posteriori validation (Weir, 2005). As mentioned before, early views of validity focused on the psychometric qualities of tests and viewed a posteriori statistical analyses as the basic indicators of validity. To quote Weir (2005, p.17) "the concern was much more with a posteriori relationship between the test and psychological abilities, traits, constructs it had measured than with a *priori* investigation of what *should* be elicited by the test

before its actual administration." Weir (2005, p.18) calls this a 'suck-it-and-see approach' and says that it is problematic to use a test before clearly identifying what we are actually measuring. According to Weir (2005), before developing a test, test developers first need to define the construct they are interested in measuring and the contexts in which the construct is elicited. In other words, test development should start with an investigation of a priori validation evidence (i.e. context validity and cognitive validity).

Context validity aspect in SFR will be explained in the next part with special emphasis on linguistic demands as this study will specifically focus on linguistic task demands as a part of context validity.

2.5.1 Context validity

Weir (2005, p.19) defines context validity as "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." As such, context validity is similar to Bachman and Palmer's (1996) situational authenticity, which is the degree to which a language test resembles the target language use (TLU) situation. Bachman and Palmer (1996) define (TLU) domain as "a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize" (p.44). Therefore, it is necessary to identify the target reading activities in terms of their contextually relevant features and try to reflect these features in the test tasks. Otherwise, it would be 'imprudent to make statements about a candidate's ability to function in typical conditions in his or her future target situation' (Khalifa & Weir, 2009, p.81).

From this perspective, one can gather context validity evidence for the use of a test of English for Academic Purposes (EAP) through exploring the similarity of the test tasks to the tasks in the TLU domain, which, in this case, is the academic language use context. The more similarity there is the more confident or truthful the test developers' inferences about a test taker's ability to function in an academic context can be. According to Weir (2005), this kind of a similarity can serve as useful validity evidence because it reduces construct irrelevant difficulty or easiness, and it increases construct representation. For instance, if the reading test tasks of an EAP proficiency test resemble the reading tasks in the TLU domain, then we can presumably say that those reading test tasks are representative of the construct in question. Without context validity, a test would inevitably lead to construct irrelevant variance, and thus the interpretations based on its results would lose their meaning. To continue with the example of the EAP proficiency reading test tasks, if the reading test tasks are considerably easier than the TLU domain tasks, then high scores cannot be assumed to guarantee that the test takers will perform effectively in reading tasks in the TLU context.

It seems obvious that one needs to generate context validity evidence for the validation of interpretations based on an exam. However, it is also important that such evidence be clear and detailed. As Bachman and Palmer (1996, p.142) put it: "If the information regarding the TLU domain is not clear and detailed, you cannot evaluate the degree of correspondence between the TLU domain and test task." In other words, there is a need to clearly specify the contextual demands of the TLU domain in a detailed way and use this information to generate clear and detailed test specifications. Similarly, Weir (2005, p.14) emphasizes the necessity of clear and explicit task specifications or descriptions that define 'both the cognitive and

linguistic abilities involved in activities in the language use domain of interest, as well as the context in which these abilities are performed (theory-based validity and context validity)'.

Another point made by Weir (1993, as cited in Weir, 2005) is that the context of an exam must be regarded by candidates and experts as a suitable one for the assessment of the language abilities of interest. In the same vein, Bachman and Palmer (1996) claim that effective representation of the TLU domain influences the reaction of test takers to the exam. In other words, test takers would not be motivated to take an exam that does not resemble the TLU context, and this would negatively affect the truthfulness of the interpretations based on that exam.

On the other hand, providing context validity evidence is not an easy task because of the limitations that the testing situation imposes. Green et al. (2010) say that achieving full situational authenticity is not realistic because of the contextual constraints that testing situations must have (e.g. time available to complete a task). Weir (2005, p.56) also admits the difficulty of achieving complete context validity saying: "Full authenticity of setting is obviously not attainable in the language test, but the settings selected for testing should be made as realistic as possible in terms of as many criterial contextual features as possible."

Khalifa and Weir (2009) suggest three kinds of context validity evidence in their framework: setting of the task, the administrative setting, and the linguistic demands of the task (see Figure 2).

The first category, task setting, involves collection of evidence on whether (1) the test format affects test performances, (2) the weighting of the tasks are justified, (3) the test takers know how they will be judged, (4) tasks or items are justifiably

CONTEXT VALIDITY	
SETTING: TASK - Response method - Weighting - Knowledge of criteria - Order of items - Channel of presentation - Text length - Time constraints SETTING: ADMINISTRATION - Physical conditions - Uniformity of administration - Security	LINGUISTIC DEMANDS TASK: TASK INPUT AND OUTPUT - Overall text purpose - Reader-writer relationship - Discourse mode - Functional resources - Grammatical resources - Lexical resources - Nature of information - Content knowledge

Figure 2 Context validity (Khalifa & Weir, 2009)

ordered, (5) the channel of presentation (computerized text versus script on paper, or plan text versus tables or graphs etc.) is justifiable, (6) the length of reading task(s) is suitable for the assessment of operations intended to measure, and (7) timing for each part of the test is appropriate. There is notable amount of research showing that the abovementioned task characteristics are likely to influence test takers' exam strategies and performance and thus the score interpretations based on them. That is why test writers need to investigate their effects and control these characteristics to make sure the tests have the desired quality (Bachman, 1996).

The second category, administration, involves whether (1) the physical conditions of the test are satisfactory (in terms of noise, seating etc.), (2) the administration of the test is uniform (or standard across different administrations), and (3) the test is secure (and kept in a safe place until it is given).

The last category, linguistic task demands, will be explained in more detail in the next part as this study focuses specifically on linguistic task demands.

2.5.2 Linguistic task demands

It is well established that linguistic demands of a reading text affect the comprehensibility of it for the readers. That is why Khalifa and Weir (2009) state that the linguistic demands imposed by the reading tasks in the future target situation should first be specified and detailed in the *priori* validation process. Only then can the test developers design test tasks that are as similar as possible to real-life language use situation in terms of linguistic demands, and only then can score interpretations be meaningfully generalizable to the real-life language use situation.

2.5.2.1 Overall text purpose

It is evident that the purpose of a writer in writing a text influences the way he or she writes in terms of various factors such as organizational style. According to Khalifa and Weir (2009) the overall purpose of a text is a factor that affects reading comprehension and therefore, it needs to be taken into account in the selection of reading test tasks. They propose Weigle's (2002) categorization of writer intentions. Some of the most dominant writer intentions listed by Weigle (2002, as cited in Khalifa & Weir, 2009) are referential (intended to inform), conative (intended to persuade), emotive (intended to convey feelings), poetic (intended to entertain), and phatic (intended to keep in touch)

It is common knowledge that academic texts usually aim to inform the reader on a point or persuade them. Consequently, university texts are highly informational and non-narrative in nature (Amjad & Shakir, 2014; Biber, 1989; Biber et al., 2002). Therefore, in an EAP reading test one might naturally expect to see non-narrative, informative and persuasive texts for the sake of context validity.

2.5.2.2 Writer-reader relationship

Writers of texts mostly write for an anticipated readership, and the characteristics of the anticipated audience are likely to affect the way writers construct their texts. According to Hyland (2002, as cited in Khalifa & Weir, 2009), in order for a writer to create the desired impact on the target reader, he or she has to understand the capacity of the readers to comprehend the text and write accordingly. In cases where the writer believes that the intended audience is familiar with the topic or concepts in the text (e.g. they are experts in the field), a big amount of topic specific background knowledge on the part of the reader might be essential to comprehend the text.

The relationship between the writer and the intended readership affects the writer's choice of vocabulary and grammar as well. Khalifa and Weir (2009) say that texts written for experts tend to include complex grammar and topic specific vocabulary that is low in frequency, which make the text difficult for readers who do not have the shared linguistic and content knowledge of the intended readership. This implies that in deciding the appropriateness of a reading text for a test of EAP, the test developers need to consider the characteristics of the intended readership (i.e. university students) just like the writers of university course books or academic text books do. However, it should also be kept in mind that the audiences of texts usually come from different backgrounds; consequently, it is not easy to make assumptions about them. Khalifa and Weir (2009) state that the age ranges and other available information about the test takers' cultural and educational background can guide test developers in choosing texts. In their analysis of Cambridge ESOL exams, Khalifa and Weir (2009) found that the reading texts in C1 and C2 level exams (i.e. CAE and CPE) require a certain amount of world knowledge and education. They also say that

these exams are aimed mostly at a young age group, so the test designers need to select appealing and accessible topics for young people.

2.5.2.3 Discourse Mode

Weir (2005) says that there is a connection between discourse mode of a reading text and the operations that readers perform. According to Alderson (2000), readers' knowledge regarding the organization of a text helps them while reading because they have an idea about where to look for a specific kind of information and how information and content changes are signaled. Alderson (2000) also emphasizes the role of discourse in reading comprehension and assessment saying "It could be that what causes difficulty in texts is less the actual content than the way the text is written" (p. 63).

There is considerable amount of research looking into the role of discourse in reading. In their study on the effect of text discourse on information recall, Meyer and Freedle (1984) found that discourse types of causation, comparison and problem-solution were found to promote better information recall than description. According to Alderson (2000), there is general agreement that narrative texts are easier to process than expository texts, probably because narratives involve a wider range of relationships between text units and also induce visualization of the written material. In an effort to analyze the discourse features of university texts Biber et al. (2002) used Biber's (1988) tagger for the multidimensional functional analysis (MAT) and found that written registers (text books, course packs etc.) at American universities shared certain discourse features such as being heavily informational and non-narrative. Weir (2005) points out that the test writer needs to check whether the

discourse mode of a reading text is appropriate for the context that the test is designed to represent.

2.5.2.4 Grammatical resources

It is widely accepted that texts with complex grammatical structures tend to be more difficult to process by the reader compared to texts with simpler grammatical structures. Grammatical complexity is usually connected to several factors such as the length of sentences, the number of modifiers per noun phrase, and the number of words before the main verb (McNamara, Graesser, McCarthy, & Cai, 2014, p.70). It is common sense that as sentences get longer, they seem to contain more of these features and thus they become grammatically more complex. That is the reason why many readability formulas use sentence length as an indicator of overall textual difficulty (e.g. Flesch-Kincaid Readability, Lexile Framework etc.). Alderson (2000) posits that the ability to parse sentences into their syntactic structures is a significant factor affecting reading comprehension. According to a study by Shiotsu and Weir (2007), syntactic knowledge is found to be a strong predictor of performance in a reading comprehension test. Alderson (1993) found that grammar test performances of students correlated highly with students' IELTS reading scores. Similarly, in their analysis of Cambridge ESOL exams, Khalifa and Weir (2009) found that grammatical complexity of reading test tasks increase as the level of exams increase.

In short, structural difficulty seems to be an important variable that might affect reading comprehension, and as such it needs to be a significant consideration in the selection and development of reading test tasks (Green et al., 2010).

2.5.2.5 Lexical Resources

Alderson (2000) states that knowledge of vocabulary has long been recognized as a crucial factor (sometimes the most crucial factor) in reading comprehension as many studies show high correlations between the readers' vocabulary knowledge and reading comprehension scores. According to Perfetti's (1985) verbal efficiency theory, identification of words through phonological recognition and meaning retrieval lies at the core of the reading activity, and the reader's capacity to process words influences comprehension. Readers spend more time when they encounter low frequency vocabulary while reading (Brysbaert & Cortese 2011). It is evident that encountering unknown words reduce readers' comprehension and negatively affect their motivation. It is suggested that in order for a reader to comprehend a text adequately and be able to predict the meaning of unknown vocabulary, he or she needs to know at least 95% of the words in the text (Hu & Nation, 2000; Laufer, 1989; Laufer & Kalovski, 2010). Lexical difficulty of texts is usually related to length of words, frequency and range of words, and lexical diversity in the text. Weir (2005) suggests that texts that contain lower frequency words tend to be more difficult compared to texts with higher frequency words. According to Cobb (2003), words that fall outside the list of most frequent 15000 words in British National Corpora (BNC) are either proper names or too technical words that increase the lexical complexity of a text for readers. "For higher-level English for Specific Purposes (ESP) students we need to examine whether the lexical range is appropriate in terms of common core, technical and sub-technical vocabulary" (Weir, 2005, p.78).

In brief, lexical suitability of reading texts to be used in tests is an important consideration for developers of reading tests (Khalifa & Weir, 2009).

2.5.2.6 Nature of information

This category relates to the level of abstractness in a reading text. Weir (2005, p.74) says that "whether the information in a text is abstract (ethics, love, etc.) or concrete (the objects in a room, for example) is relevant to the appropriateness of the test." Khalifa and Weir (2009, p. 137) comment: "It seems likely that concrete language is easier to process because it can draw upon the cognitive operations of both verbal and nonverbal (imagery) systems. In contrast, abstract language is restricted to the verbal system." According to Weir (2005) both abstract and concrete information can be found in the same text, so the test writer needs to ask oneself the question whether the nature of information is appropriate for the target language situation requirements of the candidates. According to Green et al. (2010), most of the academic texts usually deal with abstract ideas. In this respect, one can expect to see more abstract texts in the case of an EAP reading test.

2.5.2.7 Content knowledge

Khalifa and Weir (2009) point out that the interaction between the content of a text and the reader's background knowledge influences the way the reader handles the text. The relationship between the two is a symbiotic one as the content familiarity of a text depends on the reader's existing knowledge on the text topic. Weir (2005, p. 75) argues: "A text should not be so arcane or so unfamiliar as to make it incapable of being mapped onto the reader's existing schemata." Emphasizing the powerful effect of topic on readers, Alderson (2000) claim that topic unfamiliarity is undesirable in reading tests. However, he also cautions that texts that are too familiar may help readers answer the questions without depending on the text. This requires

the test developer to select or develop texts of appropriate familiarity level since unfamiliar topics might be perceived unsuitable and unfeasible while topics that are too familiar might help test takers answer questions just by using their world knowledge. In a study which investigated the influence of topic familiarity on test performance Salager-Meyer (1991, as cited in Alderson, 2000) found that topic familiarity has a stronger influence on test performance than text structure. Khalifa (1997, as cited in Khalifa & Weir, 2009) also concluded that familiarity is a strong predictor of reading comprehension scores. However, content knowledge does not only involve topic familiarity. Khalifa and Weir (2009) recommend test developers to avoid topics that might offend or upset some test takers (e.g. war, politics, religion, death etc.) or favor a certain gender or age group.

According to Weir (2005), in the EAP setting, fields such as health, education, habits, ecology etc. can be used in order to minimize bias favoring a certain group of candidates. Empirical studies in general point out that non-specialist texts on art, humanities and social sciences seem to be easier than scientific texts (Alderson, 2000). Based on similar previous findings, the writers of IELTS claim that they try to refrain from using texts that are too culturally specific or too technical for general readers.

To recapitulate what has been said with respect to linguistic task demands, there is consensual agreement in the literature that the linguistic task demands of texts listed in SFR affect the comprehension of a text. Therefore, test designers need to take into account these factors in the development of reading test tasks. According to Khalifa and Weir (2009), provision of evidence on the similarity of these features in test tasks and in TLU domain tasks can serve as context based validity evidence supporting the adequacy of score based interpretations.

2.6 Collecting context validity evidence

It is obvious that test writers need to identify the linguistic characteristics of the future TLU context and try to produce test tasks that adequately represent that context. Now that we have identified which factors to focus on, one question remains: How can one collect evidence on these linguistic demands that determine the complexity of difficulty of the reading texts? Khalifa and Weir (2009) remind that text complexity is the result of the interaction of different factors that can compensate for each other. To illustrate, even though longer sentences are often thought be more difficult than shorter ones, use of colloquial language and infrequent vocabulary might make short sentences much more difficult to understand. Similarly, Alderson (2000) notes that a reader's familiarity with a text topic can compensate for his or her lack of grammatical knowledge while reading the text. Alderson (2000) cautions that test writers need to beware of this interaction between text features and avoid simplistic approaches in developing test tasks. This implies that the test developer needs to collect evidence on multiple linguistic features instead of focusing only on one factor. Besides, as mentioned before, the evidence collected must be detailed and clear so that task specifications can be developed using such evidence because clear specifications ensure consistency of tests across administrations of different versions.

For the collection of context validity evidence Weir (2005) points to three directions*:* reviewing the literature relevant to the area, document analysis and interviews or questionnaires to get stakeholders' opinions. As for literature review, he recommends test designers to find out about how the construct intended to measure is defined by researchers and what factors are found to affect that construct. As in the case of reading assessment for instance, there is considerable amount of

research on the nature of reading and factors that seem to influence reading comprehension, some of which are mentioned above. As for document analysis, Weir (2005) suggests that data can be collected from textbooks, curriculum, official syllabuses and existing exams. For the development of an EAP reading test, for example, analyzing the task demands of TLU domain course books or existing EAP reading tests can provide contextually relevant data to the test writer. Finally, getting the opinions of stakeholders through interviews or questionnaires helps test developer better understand the TLU domain characteristics.

For the purposes of this study, all the three directions are taken. An analysis of the literature revealed what contextually relevant factors to focus on. These are already outlined above under Khalifa and Weir's (2009) SFR. For document analysis, two ways of measuring textual complexity will be adopted. First, opinions of expert judges, who comprise a part of the stakeholder population, will be referred to on deciding what kinds of textual features are relevant to the context. For this, following Bachman and Palmer's (1996, p.113) suggestion, the experts in the TLU domain (i.e. instructors at university in the case of this study) will be asked to evaluate various texts with different features in terms of different textual features and their relevance for the TLU context population. Second, automated analysis tools will be used to analyze existing documents in the TLU domain. To this end, the documents in the TLU domain (university course book texts) will be analyzed using a variety of automated textual analysis tools. At this point, it is necessary to discuss in detail how to utilize expert judgments and automated textual analysis tools to generate context validity evidence. The next part will focus on generating context validity evidence through automated tools and expert judgment.

2.6.1 Collecting context validity evidence through automated tools

Document analysis of the TLU situation with the aim of developing reading test specifications should obviously involve measurement of the textual complexity of texts in that domain (Bachman and Palmer, 1996; Weir, 2005). However, defining the complexity or difficulty of a text is not a very simple task as it involves a multitude of text and reader related factors mentioned above. As such, measuring textual difficulty has always been a major issue for the researchers in the field of language testing. Many researchers (e.g. Crossley, Greenfield, & McNamara, 2008; Dufty, Graesser, Louwerse, & McNamara, 2006; Fulcher, 1997) emphasize the important role of text difficulty in choosing texts for inclusion in exams. As mentioned before, a text at an inappropriate difficulty level might yield irrelevantly low or high scores, which threaten the accuracy of interpretations based on them.

Traditionally, expert judgments of test writers have been used in deciding the appropriateness of texts for inclusion in exams. However, partly because of the subjectivity involved in the use of expert judgments, many methods or formulas have also been developed to determine the textual difficulty of texts. According to Benjamin (2012) more than 200 textual readability formulas were developed by the 1980s. Considering the recent technological advances, one can confidently assume that a number of other formulas or methods measuring textual difficulty must have been developed since then. Thanks to the advances in computational linguistics and corpus studies, easily available computer software can instantly report textual complexity statistics. One advantage of these tools is that they can deal with big amounts of data. Additionally, they are much faster and thus more efficient compared with human raters. For example, as a recent tool developed for analyzing the lexical complexity of texts TAALES (Kyle & Crossley, 2015) presents results for more than

130 indices. As another tool developed to analyze texts at multiple levels of language and discourse, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) offers more than 100 indices. This richness presents a question for the practically oriented test writer: How can one decide which textual analysis tools or indices to use considering the abundance of available tools and the huge amount of information they provide? For this, one needs to understand different functions of these tools and the theories behind them. Benjamin (2012) identifies two methods that have inspired researchers in the analysis of textual complexity: traditional-style methods and methods inspired by advances in cognitive theory. The following part will discuss these two methods.

2.6.1.1 Traditional-style methods of measuring text complexity

Traditional-style methods are developed based on the assumption that longer sentences and longer words are more difficult to process and therefore are more difficult to comprehend. In other words, according to these methods text length and word length are the main factors that affect comprehension. Although they are called traditional, some traditional-style formulas, such as the Lexile Framework (Smith, 2000), Flesch Reading Ease, and Flesch and Kincaid Readability Formula have remained popular for years and are still used today (Benjamin, 2012). Many studies have found these formulas as valid indicators of text difficulty (e.g. Plakans & Bilki, 2016; Smith, 2000). However, although these formulas are commonly used, they are criticized for being unable to explain textual difficulty comprehensively (Benjamin, 2012). The obvious potential problem with these formulas is they take into account only the surface or lower level indicators in a text (e.g. sentence length, word length, word frequency etc.), but they fail to consider deeper level factors such as discourse

mode, cohesion between the sentences or the abstractness of the concepts in the text, all of which are known to affect text comprehension. To illustrate, an incoherent text full of short sentences composed of jumbled words might be rated as quite readable by these formulas. Some proponents of cognitive theory criticize these tools for disregarding the role of coherence in text comprehension (Britton & Gülgöz, 1991; McNamara et al., 2014). This brings the discussion to the second category in Benjamin's (2012) classification: methods inspired by advances in cognitive theory.

2.6.1.2 Methods inspired by advances in cognitive theory

Since Kintsch and Van Dijk's (1978) model of reading comprehension, many researchers have emphasized that factors other than surface level indicators also play a role in comprehension. They have stressed that reading comprehension is about constructing knowledge through integrating the new information in the text to the existing schemata of the readers. Accordingly, they have emphasized the importance of factors such as coherence, genre, discourse mode and topic familiarity as factors determining the difficulty of a text (e.g. Britton & Gülgöz, 1991; Graesser et al., 2004). Britton and Gülgöz's (1991) study, for instance, indicated that reading comprehension increases when texts are modified and made more coherent through the addition of connectors and clarifications. The study showed that even when traditional formulas rated the modified and original versions of texts similarly, students scored higher when they read the modified (more coherent) versions of texts.

Today, many cognition researchers consider traditional methods insufficient in explaining textual complexity (e.g. Benjamin, 2012; Crossley, Allen &

McNamara, 2011), yet these formulas are still popular as a number of studies have found them to be valid indicators of textual difficulty. On the other hand, it should also be noted that methods based on cognitive processing theories are quite promising as they take into consideration cognitive factors that are known to affect comprehension (Benjamin, 2012). As such, one can assume that utilizing traditional methods in combination with methods inspired by cognitive theory can yield more accurate results than resorting to one category alone. The findings from Dufty et al.'s (2006) study seem to support this assumption. In their study, Dufty et al. tried to examine how effective the different indices of Coh-Metrix (Graesser et al., 2004) was in terms of predicting the grade levels of 311 text books. Coh-Metrix provides indices based on both traditional approaches (e.g. Flesch-Kincaid Grade Level) and cognitively oriented approaches to reading difficulty (e.g. cohesion). It was found that a combination of text cohesion indices and Flesch-Kincaid was the best predictor explaining 68% of the variation between the books. It might be taken to imply that, a test developer needs to make use of a combination of various tools and textual difficulty indices to account for both traditional popular sources of textual difficulty (e.g. sentence length, word length, word frequency etc.) and sources of difficulty based on cognitive science (e.g. cohesion, discourse mode, genre, concreteness etc.). Test specifications developed using such detailed data would be assumed to help develop tests that have contextual relevance and representation.

2.6.1.3 Studies that utilized traditional and cognitively oriented indices

As mentioned before, many of the current automated textual analysis tools offer a wide range of indices. Some tools offer a combination of indices that are based on traditionally oriented and cognitively oriented approaches to capture not only the

surface but also the deeper level indicators of complexity. However, the big amount of data that such tools provide might leave the practically oriented test developer puzzled. Identifying the most efficient indices is of crucial importance. Several studies have been carried out to see which of the indices provided by automated tools successfully predict textual complexity. A study by Graesser, McNamara and Kulikowich (2011), for instance, aimed at exploring the main factors explaining most of the variance across texts at grade levels. Five major factors, among more than 100 measures that Coh-Metrix offers, were found to account for most of the variance among books across grade levels. These indices were word concreteness, syntactic simplicity, referential cohesion, causal cohesion, and narrativity percentile scores. In a similar study, Green, Khalifa and Weir (2013) analyzed the textual features of three suits of Cambridge exams, namely FCE (B2), CAE (C1), and CPE (C2), based on Khalifa and Weir's (2009) SFR. They used Coh-Metrix and VocabProfile (Cobb, 2003) to analyze the reading texts in these exams. The results of their multiple regression analysis showed that 11 indices were good predictors of text level. Four of these indices were lexical indices (AWL, mean frequency of content words, off list words, and Type Token Ratio for content words); two were syntactic indices (number of words per sentence and number of modifiers per noun phrase); and five were text level indices (anaphoric references, logical operators, concreteness, argument overlap, and content word overlap).

2.6.2 Collecting context validity evidence through expert judgment

Despite not being an automated tool or not having a certain formula, expert judgments have also been commonly used when generating context validity evidence. Green et al. (2010), for instance, state that expert judgment of test writers

are often used as validity evidence regarding text properties. As mentioned before, the context of an exam should be accepted by experts as an appropriate milieu for the assessment of the construct of interest. What is meant by '*expert'* is a person who has an acceptable level of training, knowledge or experience in the field of interest. As Berk (1990) puts it, the experts should be able to evaluate (a) the appropriateness of the content for intended populations, (b) the accuracy of the content and the domain structure, and (c) the representativeness of the content coverage of the relevant domain. For instance, in the case of developing a reading test for an EAP exam, the people who have training in test development, the people who teach EAP, or the people who make use of academic materials written in English for an acceptable amount of time may constitute an expert group and their opinions on the suitability of a task for an EAP reading test may serve as validity evidence. According to Weir (2005), the test developer might ask expert judges to rate the difficulty of a task. Taking the average of all the judges' ratings, the test developer can determine an overall difficulty estimate for a task. This way, it is possible to check whether the present tasks is at a desired level of difficulty or whether it is comparable to previous versions of the test etc.

The value of expert judgment, especially pooled expert judgment, is well attested in the field. Even the developers of automated textual analysis tools resort to pooled expert judgments in order to validate their own tools, despite the fact that one of their justifications for the development of their tools is to minimize subjectivity involved in human judgments. In other words, they validate their tools against human judgments. For instance, while examining the effectiveness of Coh-Metrix in measuring textual difficulty, Crossley et al. (2011) used texts that were categorized by expert judges in terms of their difficulty levels. The amount of correspondence

between the experts' categorizations and the tool's ratings was thought to reflect the effectiveness of the tool.

The developers of automated textual analysis tools also depend on corpora developed on the basis of experts' ratings. An example is word concreteness score that is provided by tools such as TAALES or Coh-Metrix. The concreteness scores produced by these tools are based on average concreteness scores derived from experts' concreteness judgments of words. What these tools do is basically take the average concreteness score of each word in a text and then determine an overall concreteness score for the whole text.

Expert judgments are also an important source of evidence for some features of texts that cannot be captured by automated tools. Graesser et al. (2011) state that automated tools cannot capture such textual characteristics as metaphors, or literary devices in texts. Similarly, Sheehan, Kostin, Futagi and Flor (2010) claim that automated tools are not sensitive to genre effects and they may underestimate the difficulty of literary texts. Nelson, Perfetti, Liben and Liben (2012) suggest that automated tools perform better when they are used to analyze informational texts rather than narrative texts. Green et al. (2010) point out that expert judgment is necessary in the evaluation of features such as topic familiarity, content knowledge, or cultural bias. As factors such as topic familiarity or culture specificity depend on the test taker population, the judgments of experts who are familiar with that population can provide valuable information.

In short, referring to expert judgment is a useful way of generating context validity evidence in that it can be used to validate the analysis results of automated tools as well as provide a means of analyzing the factors that cannot be evaluated by these tools.

2.7 Conclusion

Within the last 60 years, the concept of validity has evolved from a psychometric property of assessments with different types to a unified concept referring to the adequacy and appropriateness of the interpretations made based on scores. In cases where high stakes decisions are made based on exam scores, collecting validity evidence is regarded as a responsibility for the defensibility and justification of such decisions. Using a validation framework for the collection of validity evidence is commonly recommended as validation frameworks guide test developers in terms of what kinds of evidence to collect and how to collect it. Socio-cognitive Framework for Validating Reading Tests (SFR), which is developed by Weir and Khalifa (2009), is a recent framework that is found to be a practically applicable framework based on a sound theoretical basis. As Weir (2005) suggests, a priori validation (context validity and cognitive validity) is the first step in collecting validity evidence because it is pointless to administer a test and interpret its scores without clearly specifying the construct measured and the conditions under which the construct can be elicited. Collection of context validity evidence in order to generate descriptions of contextually relevant features is of critical significance for the generalizability of score-based interpretations to the TLU domain situations. In the case of an EAP reading test, the linguistic task demands of the TLU domain (academic domain) need to be thoroughly investigated and relevant linguistic factors need to be specified in a detailed way. For this, Weir (2005) recommends an analysis of the relevant documents in the TLU domain using textual analysis tools in addition to getting expert opinions to determine what is more relevant to the relevant context. However, given that such tools include a remarkably high number of indices and generate a multitude of scores for various aspects of textual difficulty, the practically oriented

test developersare bound to feel the need to limit the number of textual analysis tools and indices to focus on. It is known that the analysis results derived from such analysis tools are usually validated against expert judgments; however, it is usually the developers of the tools doing this in order to validate their own tools. There seems to be a lack of research that validates different indices from a variety of tools against the experts' judgments in a specific context, such as in an institution. Determining the indices that are validated by experts in a specific context and then analyzing the documents within that context seems like a reasonable attempt to come up with contextually relevant task descriptors (i.e. test specifications) that reduce construct irrelevant variance and construct underrepresentation. Besides, using dependable textual analysis indices, one can check whether the task demands of different contexts share certain similarities in terms of the linguistic demands placed on intended populations. In other words, the use of reliable textual analysis tools can offer a chance to make comparisons across different target domains that share certain characteristics.

With this in mind, the present study aims at investigating the following questions:

- Research Question 1a: What are the factors that explain the experts' judgments on the difficulty of a text?
- Research Question 1b: What are the factors that explain the experts' judgments on the suitability of a text?
- Research Question 2: Which automated tools can explain experts' judgments of a text?
- Research Question 3: Do texts from different corpuses differ from each other with respect to the features analyzed by automated tools?

- Research Question 4: What are the optimal ranges of text characteristics that will account for the texts that are gathered from three different EMI contexts?

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the methodology used in the current study is presented. Firstly, the context, the participants and the instruments used are explained. Then, the procedures and the analyses with respect to each research question are provided.

3.2 Context

As one focus of the current study involves context validation, it was essential to find a context from which to collect relevant data. During the time data was collected, the researcher worked at İstanbul Şehir University (ISU), a Turkish foundation university whose medium of instruction is English for all majors except for Turkish History and Turkish Language and Literature. For this reason, the participants were chosen among the instructors working at ISU.

At ISU, students are expected to be proficient enough in English to be able to follow the course content. To ensure that, the school requires students to prove their English competency through providing an acceptable score in TOEFL, PTE, IELTS or the school's own in-house proficiency test, STEP. Those who cannot provide any of the above evidence after registration are required to study at the School of English Preparatory Program (SEPP) until they pass one of the abovementioned exams. A big majority of students coming to ISU are not at the proficiency level the school expects; therefore, they study at SEPP. SEPP has a modular system, in which each student is placed on a level based on a placement test given at the beginning of the

year. After studying at a level for seven weeks (one module), students take a module-end test to move on to the next level. There are five levels at SEPP: Elementary (A1), Pre-Intermediate (A2), Intermediate (B1), Upper Intermediate (B2), or Pre-Faculty (B2+). Students can take STEP only after they complete the Pre-Faculty level. Those who pass STEP are allowed to start their majors in their faculties.

Although the participants of the study were selected from the context of ISU, materials from other contexts were also needed for the investigation of the third and fourth research questions that aim to investigate whether the textual demands placed on students or test takers are similar in different contexts and whether it is possible to generate optimal ranges of textual characteristics based on a comparison of corpuses. To this end, texts from two other universities and IELTS preparatory materials were also used. The universities were Boğaziçi University (BOUN), a state university in Turkey with about 16000 registered students as of 2016, and University of Bedfordshire (UBU), a public university in England with about 24000 registered students as of 2016.

It should be noted that although İstanbul Şehir University (ISU), Boğaziçi University (BOUN) and University of Bedfordshire (UB) are all English medium universities, their expectations from students with regard to their English proficiency are different. In this respect, one can assume that the contexts of these universities place different linguistic demands on students. Table 1 summarizes the minimum acceptable levels of English proficiency students at each university are required to pass.

Table 1. Minimum Scores of English Proficiency Accepted by Universities

University	TOEFL IBT	IELTS	PTE
BOUN	79	6,5	-
ISU	79	6,0	55
UB	72	6,0	51

3.3 Participants

In order to investigate the nature of experts' judgments on different characteristics of texts for research question 1a and 1b (RQ1a and RQ1b), the researcher needed an expert group that could be considered capable of evaluating the relevance of texts to a specific purpose in a certain context. As Berk (1990) states, experts need be chosen from people who can judge the (a) appropriateness , (b) accuracy, and (c) representativeness of the content coverage in relation to the domain. In the context of ISU, these experts had to be chosen from people who were familiar with the characteristics of the students at ISU and the contextual demands of the TLU domain. For this, the researcher identified two groups of people: the English instructors at SEPP and the core curriculum instructors at faculty level. About 80 instructors were contacted personally or via e-mail. They were explained the purpose of the study and asked whether they would like to participate. Of the instructors contacted, 37 SEPP instructors and 10 faculty instructors agreed to participate.

The 37 SEPP instructors who participated in the study had been teaching English at SEPP for at least one year by the time they were given the questionnaire for the present study towards the end of 2016-2017 academic year. At the time, these instructors were teaching English for 12 to 20 hours a week to a class of about 20 students at one of the five levels at ISU. At the end of each seven-week module, a teacher was assigned another class, often at another level. Therefore, in an academic year, most SEPP teachers taught a total of five different classes during the five

modules, and spent time with a total of about 100 students, who can be regarded as a representative sample of the general student population at SEPP. For the purposes of the current study, the researcher purposefully selected SEPP teachers who taught a Pre-Faculty class at least once in 2016-2017 academic year. The reason for this was that Pre-Faculty was the level where students were exposed to tasks that resemble faculty studies (e.g. reviewing a research article), and as such the Pre-faculty teachers were expected to be quite familiar with the expectations of the TLU domain as well as the student profile at SEPP.

The second group of participants was composed of 10 core curriculum instructors. As stated in ISU Core Curriculum Brochure, core curriculum was an independent program at ISU, and it was designed to "give students a holistic look on life beyond conventional academic compartmentalization, such as natural and social sciences." These courses were a group of elective courses among which all ISU students were required to select some and pass. The number of core courses to take varied depending on the major of the student, but every student at the university needed to take at least 13 core courses before they could graduate. The reason why the researcher purposefully selected core course instructors was that these instructors were familiar with the general student body of ISU as they saw students from all departments, and their courses were not too field specific.

3.4 Instruments

Different instruments were used in order to collect and analyze the data for relevant research questions. The instruments used included a questionnaire and various automated textual analysis tools.

3.4.1 The questionnaire

A questionnaire was used in order to explore the instructors' judgments on different textual characteristics (see Appendix A). The questionnaire included 10 extracts of about 200 words each (see Table 2). The first five texts were taken from the course books of five ISU core courses: Understanding Society and Culture (UNI 117), World Civilizations (UNI 221), Understanding Ethics (UNI 213), Textual Analysis and Effective Communication (UNI 123), and Understanding Science and Technology (UNI 203) respectively. Texts 6 (ICork), Text 7 (IDis) and Text 10 (IHob) were extracts taken from Official IELTS preparatory materials (IELTS 12 Academic) published by Cambridge Press. It should be noted that texts with different characteristics were included in the questionnaire in order to see the judgments of the experts on different text features. However, it was not possible to find an authentic course book or IELTS text that lacked cohesion. For this reason, the cohesion of Text 7 (IDis) was intentionally distorted by the researcher. Texts 8 (NFoo) and 9 (NPol) were newspaper extracts from CNN and BBC news web pages published in June 2017. The texts varied in terms of their linguistic and discourse characteristics. Therefore, the selection of texts represented several features to be examined by the participants and the automated textual analysis tools. However, the researcher avoided texts with extreme features (e.g. an A1 level text) to exclude texts that

Table 2. Sources of the Questionnaire Texts

Text Number	Text Source
Text 1	ISU course book for *Understanding Society and Culture* (UNI 117)
Text 2	ISU course book for *World Civilizations* (UNI 221)
Text 3	ISU course book for *Understanding Ethics* (UNI 213)
Text 4	*Of Mice and Men* (by John Steinbeck)
Text 5	ISU course book for *Understanding Science and Technology* (UNI 203)
Text 6	*IELTS 12 Academic* by Cambridge University Press
Text 7	*IELTS 12 Academic* by Cambridge University Press
Text 8	News article from CNN news website dated June 5, 2017
Text 9	News article from BBC news website dated June 8, 2017
Text 10	*IELTS 12 Academic* by Cambridge University Press

could easily be discarded by participants as being irrelevant to an academic context. On the first page of the questionnaire the participants were given some brief information regarding the purpose and the contents of the questionnaire. This was followed by individual texts and 13 questions regarding the features of each text. The first two questions were categorical in nature as the participants were asked to choose the option showing the purpose of the text (Q1) and its possible source (Q2). The following 10 questions were intended to elicit answers that would give numerically comparable information with the results of the automated tool analyses. They were in the form of 5 point Likert scale where "1" represented the easy end with respect to different linguistic and discourse features while "5" represented the difficult end. The following nine questions addressed a specific feature of the text in question. These questions were related to the target audience (Q3), grammar (Q4), vocabulary (Q5), concreteness of the information (Q6), density of the information (Q7), topic specificity (Q8), culture specificity (Q9), cohesion (Q10), and coherence (Q11) of each text. After each question, some space was provided so that the participants could write their comments. Question 12 (Q12) involved the overall difficulty judgment for the text in question, and the last question (Q13) was a Yes/No question which asked whether or not a participant viewed a certain text as a (un)suitable one for inclusion in STEP, the EAP proficiency test at ISU.

3.4.2 Automated textual analysis tools and indices

In order to compare whether there is agreement between the judgments of the experts on textual features and the results of automated textual analysis tools (RQ2), the researcher needed to analyze the 10 extracts in the questionnaire (see Appendix B for the analysis results for the questionnaire texts). Automated textual analyses tools were also used to generate data for the comparison of texts from different corpuses

(RQ3). These tools were chosen based on their availability, user friendliness, novelty and popularity. Following Dufty et al. (2006) and Benjamin (2012), the researcher included both traditional indices that focus on surface level indicators of textual complexity (e.g. sentence length, word length etc.) and cognitively based ones (e.g. concreteness of the information, coherence etc.)

One of the tools used is Multidimensional Analysis Tagger (MAT) (Nini, 2014), which is a freely available software program based on Biber's (1988) tagger for multidimensional analysis of texts in English. According to Biber (1989), although every text differs in terms of linguistic characteristics, some linguistic features systematically and frequently occur across various genres or text types, so these co-occurring features can be used to categorize texts across multiple dimensions. Biber (1988) proposes a five dimension variation model. These dimensions are (1) involved versus informational discourse (the extent to which a text has affective and interactional content or informational content), (2) narrative versus non-narrative concerns (the extent to which a text is story-like), (3) elaborated or situation dependent reference (the extent to which a text is dependent on context as in sport broadcasts or independent from context as in academic prose), (4) overt expression of persuasion (the extent to which the writer's intentions are explicitly stated), and (5) abstract versus non-abstract information (the extent to which the information in the text is abstract). Biber (1989) analyzed 481 texts from 23 different genres based on these five dimensions and identified eight common text types in English, namely (1) intimate interpersonal interaction, (2) informational interaction, (3) scientific exposition, (4) learned exposition, (5) imaginative narrative, (6) general narrative exposition, (7) situated reportage, and (8) involved persuasion. Analyzing the frequently occurring linguistic features within a text, Biber (1989) determined

which of the abovementioned types a text was closest to. It was found in the study that of the academic texts analyzed, 44% was categorized into type 3 (scientific exposition) which means these texts were informational expositions that had highly technical words and abstract language (as in a Science or Engineering text). 31% of academic prose was categorized as type 4 (learned exposition) which was very similar to scientific exposition but included less abstract and less technical content (as in a Sociology or Humanities text). 17% was categorized as type 6 (general narrative exposition) which was again a type of exposition but with frequent use of "past tenses" and "active voice" instead of "present tenses" and "passives" as in the other two types of exposition. A common way of labeling texts in academic contexts is calling them "informative" or "expository". However, although calling a text expository may help us understand what the writer's primary intention is, as Biber's (1989) study showed, exposition in itself could be grouped into three categories based on the linguistic features involved. Using Biber's tagger, a test writer may collect evidence related to overall text purpose, writer-reader relationship, discourse mode, and nature of information in Khalifa and Weir's (2009) SFR. Consequently, using the information gathered through MAT (Nini, 2014), the test writer can make more informed decisions about what type(s) of exposition best represent the TLU domain.

Coh-Metrix (Graesser et al., 2004) is another tool used in the study. Coh-Metrix can be used to analyze texts at multiple levels (e.g. word level, sentence level and discourse level). A public version of the tool (Coh-Metrix 3) is freely available online and provides more than 100 indices. Other than simple descriptive indices (e.g. text length, number of sentences, number of paragraphs etc.), Coh-Metrix offers indices of referential cohesion (e.g. noun, argument or content word overlap across

the text or between adjacent sentences) and indices of deep cohesion (e.g. incidence scores for causal words, intentional words and their ratios). These indices are thought to reflect the assumption that lies at the heart of this tool: cohesion is a very significant factor that affects reading comprehension, which is not captured by traditional readability formulas (McNamara & Graesser, 2011). Coh-Metrix also offers indices of syntactic complexity (e.g. left embeddedness, number of words per sentence, number of modifiers per noun phrase, noun phrase density etc.), lexical indices (e.g. word frequency, word concreteness, type-token ratio etc.), overall text easibility scores for syntactic simplicity, narrativity, referential and deep cohesion. Finally, in addition to the traditional readability scores of Flesch Reading Ease and Flesch and Kincaid Grade Level, the tool provides an overall readability score named Coh-Metrix L2 Readability Score, which the designers claim is more suitable to measure the difficulty of a text for L2 readers. As mentioned above, this tool provides more than 100 indices, but for feasibility Coh-Metrix indices used in this study were limited to 14 indices. These were three overall readability indices (Flesch Reading Ease, Flesch Kincaid Grade Level, and Coh-Metrix L2 Readability), five grammar indices (average number of words per sentence, average number of words per noun phrase, noun phrase density, left embeddedness, and an overall syntactic simplicity percentile score that takes into account multiple syntactic difficulty indices), three overall cohesion indices (Type token ratio and referential and deep cohesion percentile scores that take into account other atomistic cohesion index scores), CELEX word frequency for content words (CELEX CWF), Coh-Metrix concreteness percentile score and Coh-Metrix narrativity percentile score.

Coh-Metrix indices might provide data on different aspects listed in Khalifa and Weir's (2009) SFR. Cohesion indices and narrativity score can be used to collect

information on discourse mode and overall text purpose; lexical indices (CELEX CWF and Concreteness percentile) can be used to gather evidence on lexical resources, nature of information and content knowledge; and syntactic indices can provide information on grammatical resources. Finally, the overall L2 readability score, in combination with the other indices, can inform the test developer on writer-reader relationship as writers of texts adjust the grammatical, lexical and content features of their texts taking into account an assumed audience.

The Lexile Framework for Reading (Lexile) is a tool developed by Metametrics Inc. with the aim of matching learners with texts of appropriate difficulty. Lexile evaluates reader ability and textual difficulty on the same scale. Thousands of books have been evaluated and given a Lexile score, especially in the United States. After taking a reading test given by Metametrics Inc., readers are also given a Lexile score. The Lexile scores range from 0L (for the lowest level texts and readers) to 2000L (for advanced level texts and readers). Using these scores, Metametrics Inc. tries to offer a practical way to decide which books are appropriate for learners at certain levels. Lexile scores are based on traditional measures of text difficulty, namely average sentence length, word length, and word frequency. In a study by Williamson, Stenner, Sandvik, and Johnson (2016), the complexity of texts from UK universities were compared with those from American universities. The results show that averaging around 1300L, the complexity scores of texts in both contexts were very similar to each other. Checking the Lexile scores of the TLU domain texts, test developers might have an idea on the overall difficulty of the texts based on traditional, surface level indicators of difficulty.

Tool for the automatic analysis of lexical sophistication, abbreviated as TAALES, (Kyle & Crossley, 2015) is another freely available tool that was used in

the present study. As the name suggests, it is developed to analyze texts in terms of their lexical properties. One advantage of TAALES is that it can process multiple texts at once, unlike Coh-Metrix and Lexile, which can analyze only one text at a time. TAALES indices can be put into five categories: word frequency indices, range indices, n gram frequency indices, academic word list indices, indices based on psycholinguistic properties. TAALES provides results for these indices based on multiple traditional and contemporary corpuses, such as Kucera-Francis written frequency, British National Corpus (BNC) and Corpus of Contemporary American English (COCA). With the aim of validating the indices provided by TAALES, Kyle and Crossley (2015) conducted a study to see the extent to which these incides would predict the variance in the holistic human ratings of lexical proficiency based on student essays. They found that TAALES indices explained 47.5% of the variance in the holistic human ratings of lexical proficiency. They concluded that although word level frequency measures are good measures of lexical proficiency, range indices also accounted for some of the variance and thus they needed to be taken into account when designing assessment tasks. TAALES frequency, range and concreteness indices were employed in the present study. Frequency and range scores based on BNC and COCA written and spoken corpuses were checked for the questionnaire texts. Additionally, two concreteness indices (MRC and Byrsbaert content word concreteness indices) were also used. Test developers can use TAALES indices in gathering evidence for SFR context validity aspects such as lexical resources, nature of information, writer-reader relationship, and content knowledge.

3.5 Data collection

As data collection procedures were different for each research question, this section will discuss the procedures focusing on each question separately.

The first research question (RQ1) aimed to explore the nature of text difficulty and suitability from the viewpoint of experts. As such, it was formulated as: "What are the factors that explain the experts' judgments on the difficulty and suitability of a text?" The question had two sub-questions, one focusing on text difficulty (1a) and the other focusing on text suitability (1b). A questionnaire was used to collect data from the 47 instructors who agreed to participate in the study. Depending on their preference, the instructors were given either a hard copy of the questionnaire or a soft copy sent by email. The researcher explained the participants the rationale for the research. They were also told that participation was on a voluntary basis and that their names would be kept anonymous. The participants were given three weeks to complete the questionnaire so that the time pressure would be reduced. After collecting the data, the researcher computerized all the data to make it ready for analysis.

The second research question was formulated as "Which automated tools can explain experts' judgments of a text?" To explore this question, the researcher needed textual analysis scores for the texts that were comparable to the judgments of the instructors. For this, the 10 texts in the questionnaire were analyzed using the previously mentioned textual analysis tools, which are MAT (Nini, 2014), Coh-Metrix (McNamara et al., 2014), TAALES (Kyle & Crossley, 2015), and Lexile by Metametrics Inc. It is worth pointing out that be noted that although MAT and Lexile each offer one index result, TAALES and Coh-Metrix offer multiple indices that can be used. Coh-Metrix indices used for this question were (1) narrativity percentile

score, (2) average number of words per sentence, (3) syntactic simplicity percentile, (4) left embeddedness, (5) noun phrase density, (6) average number of modifiers per noun phrase, (7) concreteness percentile, (8) type token ratio, (9) CELEX content word frequency, (10) referential cohesion, (11) deep cohesion, (12) Flesch Kincaid grade level, (13) Flesch reading ease, and (14) Coh-Metrix L2 readability. TAALES indices used for this question were (1) BNC written frequency for content words, (2) BNC written range for content words, (3) BNC spoken frequency for content words, (4) BNC written frequency for academic words, (5) COCA written frequency for content words, (6) COCA written frequency for academic words, (7) COCA written range for academic words, (8) MRC concreteness score, and (9) Byrsbaert concreteness score.

The third question was formulated as "Do texts from different corpuses differ from each other with respect to the features analyzed by automated tools?" To compare texts from different contexts, the researcher first needed to identify separate corpuses each representing a context. Choosing 30 texts from each context, the researcher had a total of 120 texts for 4 corpuses. The texts were about 800 words in length each. For the university corpuses, the texts were selected from the beginning, middle and final chapters of 10 books from each university context. For the ISU corpus, the texts were extracted from 10 core course books. These were the books for UNI 102 (Critical Thinking), UNI 117 (Understanding Society and Culture), UNI 118 (Understanding Politics and Economy), UNI 123 (Textual Analysis and Effective Communication), UNI 203 (Understanding Science and Technology), UNI 204 (Understanding Nature and Knowledge), UNI 205 (Understanding Science and Environment), UNI 213 (Understanding Ethics), UNI 221 (World Civilizations and Global encounters I), and UNI 222 (World Civilizations and Global Encounters II).

The rationale behind choosing these books was that these were among the most popular core courses that were offered almost every semester and the soft versions of the course books for these 10 courses were available to the researcher at the time of the study.

For the BOUN corpus, a similar text selection procedure was followed. 10 BOUN course books were used. These were used for the following courses: HUM 102 (Cultural Encounters II), LING 101 (Introduction to Language and Linguistics), PHIL 105 (Informal Logic), PSY 101 (Introduction to Psychology), PRED 154 (Academic Orientation for Math and Science), EC 102 (Macroeconomics), TRM 104 (Environment and Tourism), HIST 106 (The Making of the Modern World), ED 131 (Comparative Legal Structures), and ED 104 (Social Foundations of Education). These were again frequently elected courses and the texts extracted from the related text books represented a good selection of first year text books from the area of social sciences.

As for the UB corpus, the researcher again used a readily available corpus. It was part of the same corpus used in Green et al.'s (2010) study in which the UB corpus was compared against a corpus of IELTS texts. The 10 texts came from first year books on Business, Criminology, Human Resources and Management, Interactive Children, Java, Law, Multimedia, Nature of Psychology and Principles of Marketing.

Finally, for the IELTS corpus, the researcher extracted 30 texts from five recent official IELTS preparation books published by Cambridge University Press between 2011 and 2017. The books, IELTS 8, IELTS 9, IELTS 10, IELTS 11, and IELTS 12, contain authentic texts from past IELTS exams. The researcher took six reading texts from each book to get a total of 30 texts representing the IELTS corpus.

All four corpuses were analyzed using the same MAT, Lexile, TAALES, and Coh-Metrix indices used for the previous research question.

The fourth research question was formulated as "What are the optimal ranges of text characteristics that will account for the texts that are gathered from three different EMI contexts?" As the three universities in question, namely ISU, BOUN, and UB, are all English-medium institutions (EMI) providing higher education, one might expect to find a lot of commonalities between the texts used in these contexts. However, judging from the different scores these universities require from students in exams such as TOEFL and IELTS, one can also assume that their texts may vary in terms of complexity. The corpuses from three universities for the RQ3 were also used for the RQ4 to explore the optimal ranges of text characteristics representing all three university corpuses.

3.6 Analyses

This section presents the analyses that were conducted to address each research question. The analyses conducted for each research question are presented separately.

3.6.1 Analyses for RQ1a and RQ1b

The research questions 1a and 1b were as follows:

- Research Question 1a: What are the factors that explain the experts' judgments on the difficulty of a text?
- Research Question 1b: What are the factors that explain the experts' judgments on the suitability of a text?

To answer these questions, the instructors' answers in the questionnaire are used. As mentioned before, each of the first 11 questions in the questionnaire focused on a specific characteristic of the text in question, while Q12 asked instructors to rate the overall difficulty of that text and Q13 asked whether they found the text suitable for STEP, EAP proficiency test at ISU.

For RQ1a, correlation analysis was first conducted to see the correlation of expert judgments on atomistic text features (Q3-Q11) to the overall difficulty judgments (Q12). Following the correlation analysis, a stepwise multiple regression analysis was used to see the predictive strength of the experts' judgments on atomistic features of the texts in explaining their overall difficulty judgments.

For RQ1b, the aim was to explore which atomistic judgments of the experts (Q3-Q11) significantly predicted whether a text would be regarded as suitable or not by the experts (Q13). However, due to the fact that Q13 was a Yes/No question, the answers to this question were categorical rather than numerical in nature. For this reason, a logistic regression analysis was conducted. In addition to the logistic regression, a qualitative analysis of experts' comments for text suitability question was done in order to find whether there were certain text features that seemed to drive experts' perceptions on text suitability.

3.6.2 Analyses for RQ2

Research Question 2 was formulated as: Which automated tools can explain experts' judgments of a text?

This question sought to investigate whether automated textual analysis indices correlated with the average expert judgments (i.e. pooled expert judgments). For this, two separate analyses were conducted. Firstly, for the categorical data from

Q1 in the questionnaire (i.e. the purpose of the text), a comparison was made between the expert judgments and two automated tools, namely MAT and Coh-Metrix narrativity percentile scores. Next, for the numerical data from Q3 to Q11, a correlation analysis was conducted to see the correlation between average expert judgments (pooled expert judgments) on each textual feature and the previously mentioned indices mentioned in this chapter. The mean scores for each textual feature given by the experts were correlated with the scores of the textual analysis tools on the same feature.

3.6.3 Analyses for RQ3

Research Question 3 was formulated as: Do texts from different corpuses differ from each other with respect to the features analyzed by automated tools?

This question aimed to investigate whether corpuses from İstanbul Şehir University (ISU), Boğaziçi University (BOUN), University of Bedfordshire (UB) and IELTS contexts differed from each other with respect to different textual features. First, the corpuses were analyzed through MAT to explore the types of texts in terms of their discourse modes. Then, the corpuses were compared based on their scores on textual analysis indices. A one-way between groups ANOVA was used to compare the 120 texts from four corpuses in terms of their textual features based on the scores produced by textual analysis tools.

3.6.4 Analyses for RQ4

Research Question 4 was formulated as: What are the optimal ranges of text characteristics that will account for the texts that are gathered from three different EMI contexts?

This question aimed to offer some general textual characteristics that seemed to represent the general characteristics of the corpuses from the three universities. 90 texts from three corpuses were combined as if it is one single corpus. Then using descriptive statistics, the ranges of textual characteristics were determined through automated textual analysis tools. Next, outliers were removed and using the Empirical Rule, which is also known as the '68-95-99 Rule', the range of the values falling within one standard deviation of the mean for each index was calculated to offer ranges that are representative of the three university corpuses.

3.7 Conclusion

This chapter explained the contexts, participants, data collection instruments and the analyses conducted to answer the research questions. Table 3 summarizes the research questions, the instruments, and the analyses to address each question. The next chapter deals with the results for the research questions and their discussions.

Table 3. Instruments and Analyses Used for Each Research Question

Research Questions	Instruments	Analyses
1a. What are the factors that explain the experts' judgments on the difficulty of a text?	Questionnaire	Correlation Analysis Stepwise Multiple Regression
1b. What are the factors that explain the experts' judgments on the suitability of a text?	Questionnaire	Logistic Regression Qualitative Analysis of Experts' Comments
2. Which automated tools can explain experts' judgments of a text?	Questionnaire Automated Textual Analysis Tools	Qualitative Analysis of Experts' Judgments Correlation analysis
3. Do texts from different corpuses differ from each other with respect to the features analyzed by automated tools?	Automated Textual Analysis Tools	One-way Between Groups ANOVA
4. What are the optimal ranges of text characteristics that will account for the texts that are gathered from three different EMI contexts?	Automated Textual Analysis Tools	Descriptive Statistics

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results of the quantitative and qualitative analyses explained in the previous chapter and the discussions based on the results. Following the presentation of the results with respect to each question, a relevant discussion is given. The chapter is concluded with a general discussion of the findings.

4.2 The results and discussion for RQ1a

This question sought to explore which judgment(s) on the text qualities made by the experts predicted the overall difficulty judgments of a text. As mentioned before, in an attempt to answer this question, a questionnaire consisting of 10 texts of about 200 words each was given to 47 experts. Table 4 presents the sources of the questioınnaire texts in addition to the word count and assigned texts codes for each text. The experts were asked to evaluate each text in terms of different linguistic and discourse features by answering 11 questions, which were developed based on SFR by Weir and Khalifa (2009). The questionnaire had two more questions (Q12 and Q13) which required the experts to evaluate the overall difficulty of each text and the suitability of each text for the proficiency test at ISU (STEP).

The first two questions were categorical in nature as the experts chose from the options as to the purpose (Q1) and source (Q2) of each text. For the following 10 questions, namely expertise required to read the text (Q3), grammatical difficulty of the text (Q4), lexical difficulty of the text (Q5), concreteness of the vocabulary in the

Table 4. Texts Used in the Questionnaire

Text Number	Text Code	Text Source	Number of Words
Text 1	Soc	ISU course book for *Understanding Society and Culture* (UNI 117)	206
Text 2	His	ISU course book for *World Civilizations* (UNI 221)	207
Text 3	Eth	ISU course book for *Understanding Ethics* (UNI 213)	203
Text 4	Nov	*Of Mice and Men* by John Steinbeck	210
Text 5	Sci	ISU course book for *Understanding Science and Technology* (UNI 203)	216
Text 6	ICor	*IELTS 12 Academic* by Cambridge University Press	213
Text 7	IDis	*IELTS 12 Academic* by Cambridge University Press	205
Text 8	NFoo	News article from CNN news website dated June 5, 2017	225
Text 9	NPol	News article from BBC news website dated June 8, 2017	210
Text 10	IHob	*IELTS 12 Academic* by Cambridge University Press	213

text (Q6), density of information (Q7), topic specificity (Q8), culture specificity (Q9), connection between sentences (Q10), the flow of the text (Q11) and the overall difficulty of the text (Q12), the experts were asked to rate each text out of five. The last question (Q13) was related to the suitability of a text for inclusion in STEP and it was a Yes/No question and the answers were coded 1 and 0 respectively. Therefore, a suitability ratio score between 0 and 1 for this question was calculated for each text based on the answers of the expert judges (see Appendix C for a summary of expert judgments of questionnaire texts).

As mentioned in the previous chapter, each text was given an overall difficulty score out of five for Q12. The median and mean scores for overall difficulty were calculated to see whether it was possible to divide the texts into two groups as relatively difficult and easy texts. The median score was 3.00 for the difficulty judgments and the mean was 2.87. There was an evident break among the texts in terms of their mean scores. Four of the texts had an overall difficulty score of

2.53 and below while the other six texts had overall difficulty scores over 3.06. Accordingly, Texts 4, 2, 7, 3, 6, and 5 were labeled as relatively difficult with average difficulty scores of 3.83, 3.55, 3.16, 3.11, 3.09, and 3.06 respectively. The remaining four texts (Texts 8, 1, 9, and 10) were labeled as easy texts with average difficulty scores of 2.53, 2.28, 2.11, and 1.98 respectively.

In order to see the predictive strength of the experts' judgments on atomistic features of the texts in explaining their overall difficulty judgments, a correlation analysis was followed by a stepwise multiple regression analysis. In other words, two analyses were conducted to investigate which of the expert judgments from Q3 to Q11 seemed to explain the experts' judgments for Q12. The correlation of judgments on each feature to the overall difficulty judgments is presented in Table 5 (see Appendix D for the correlations between all the judgments).

All the atomistic feature judgments of experts were found to correlate significantly with their judgments of overall difficulty ($p < .001$). As can be seen in the Table 5, the experts' overall difficulty judgments had the highest correlation with their vocabulary judgments, which was followed by grammar, topic specificity and so on. The nine features that correlated significantly with the overall difficulty judgments of experts were entered into a stepwise multiple regression analysis to see their relative predictive strengths. The assumptions of regression analysis were checked. It was

Table 5. Correlation of Judgments on Atomistic Features to Overall Difficulty

Feature	r value
Q5 (Vocabulary)	.71
Q4 (Grammar)	.60
Q8 (Topic Specificity)	.55
Q7 (Density of Information)	.48
Q11 (Flow of Ideas)	.45
Q10 (Connection of Sentences)	.44
Q9 (Culture Specificity)	.41
Q3 (Expertise Required)	.36
Q6 (Concreteness)	.36

found that the assumption of multicollinearity was violated as two of the scores (Q10 and Q11) had a correlation that was higher than .70. Therefore, one of the questions (Q10) was not entered into the analysis. The other assumptions of regression analysis were met.

Table 6 illustrates the results of the stepwise multiple regression analysis carried out to find the best predictors of overall difficulty judgments. As can be seen in Table 6, the experts' judgments on vocabulary, flow of the text, topic specificity, grammar, and information density were the factors that significantly predicted judgments on overall difficulty of texts, $R^2 = .67$, F (5,451)= 182.90, $p < .001$. Of these five features, vocabulary judgments accounted for the biggest variance $R^2 = .50$, F (1,455) = 445.93, $p < .01$. Taken together, vocabulary and flow of ideas explained 62% of the variance in overall difficulty judgments. The contributions of topic specificity, grammar and information density to the explained total variance were relatively small when compared with those of vocabulary and flow.

The results suggest that when experts judge the overall difficulty of a text, vocabulary difficulty, flow of ideas, specificity of the topic, grammatical difficulty and the density of the information presented in the text were found to be the factors that significantly predicted experts' overall difficulty judgments. The results also indicated that, of the five predictive features, vocabulary difficulty of a text was the most significant indicator of the experts' holistic difficulty judgments explaining50% of the variance. It is worth pointing out that although judgments on grammar had the second highest correlation with overall difficulty judgments, the regression analysis showed that the contribution of grammar judgments to the model in terms of its predictive power was quite low. This is probably due to the high correlation between the vocabulary and grammar judgments (r .68, p < .001) (see Appendix D). In other

Table 6. Stepwise Multiple Regression Analysis Results

Model	Variable added	R^2	df	F	Sig.
1	Q5 (Vocabulary)	.50	455	446	.000
2	Q11 (Flow of Ideas)	.62	454	373	.000
3	Q8 (Topic Specificity)	.65	453	277	.000
4	Q4 (Grammar)	.66	452	222	.000
5	Q7 (Information Density)	.67	451	183	.000

words, most of the variance in overall difficulty explained by the grammar judgments was already accounted for by the vocabulary judgments. This might mean that the experts' judgments on grammar and vocabulary are highly dependent on each other. Such a finding is in line with the claims of Alderson and Kremmel (2013) and Römer (2009) who assert that there is an overlap between syntactic and lexical knowledge by their nature and that they are not clearly separable characteristics. This finding also supports Sinclair (2004), who suggests 'lexicogrammar' as a unitary concept rather than treating vocabulary and grammar separately.

The finding that experts inherently rated vocabulary difficulty and overall difficulty of a text similarly seems to support the claims of researchers who previously highlighted the importance of vocabulary knowledge in the comprehensibility and thus the difficulty of reading texts. For instance, Alderson (2000) states that vocabulary knowledge is often found to have a strong correlation with reading comprehension; Laufer and Kalovski (2010) and Hu and Nation (2000) offer minimum percentages of vocabulary knowledge required to gain an adequate understanding of a text, and Perfetti (2007) and Byrsbaert and Cortese (2011) highlight the influence of vocabulary length and frequency on reading time and thus reading comprehension.

The descriptive summary of the experts' judgments (see Appendix C) also supports the regression analysis results. Of the six texts that were found to be

relatively difficult with overall difficulty scores of 3.06 or above, five (Texts 2,3,4,5 and 6) were ranked the highest in vocabulary difficulty. That is to say, the texts that were found to be the hardest in terms of vocabulary were also found to be the hardest in overall difficulty. Only one of the texts did not fit into this generalization. The Text 7 (IDis) was judged to be a difficult text overall although it was not rated difficult with regards to its vocabulary. As mentioned in the previous chapter, this text was intentionally distorted by the researcher in terms of coherence. The connectors between sentences and some words were intentionally changed to make it harder for the reader to create a logical picture of what the text says. It is very likely that these changes caused the text to be rated the most difficult text in terms of connection between sentences (Q10) and the flow of the text (Q11). As the most incoherent text in the questionnaire, it was also found to be very difficult. This again supports the regression results, according to which the flow of the text (Q11) was the second best predictor of overall difficulty. An important point related to coherence is that although text coherence did not necessarily cause a text to be judged as easy (e.g. Text 2), lack of coherence caused a text to be perceived as difficult overall (e.g. Text 7).

Another point worth mentioning is related to the judgments on connection between sentences (Q10) and flow of ideas in the text (Q11). Q10 and Q11 were intended to see the experts' judgments on cohesion and coherence of a text respectively. Because they correlated significantly with each other (r .88, $p < .001$) (see Appendix D), Q10 did not enter the multiple regression analysis. This is probably because Q11 already accounted for most, if not all, of the variance accounted for by Q10 due to the high level of correlation between them. In this respect, one can assume that cohesion and coherence of a text are features that are

dependent on each other and are viewed very similarly by the experts. This might be taken to suggest that some text features are hard to distinguish by human raters; therefore, it may be futile to ask questions on such indistinguishable features in the same questionnaire.

In short, the distinctive features of the texts that were judged to be difficult were either their vocabulary difficulty or incoherence. Based on the results, one can presumably assert that the expert judges had a tendency to equate text difficulty with vocabulary difficulty. If the vocabulary of a text is perceived to be difficult, then that text is very likely to be perceived as difficult overall. The results also suggest that a flaw in the logical progression of ideas in a text might significantly increase the overall difficulty of a text for the reader even though the text is not difficult in terms of other linguistic features, such as vocabulary.

4.3 The results and discussion for RQ1b

This question aimed to investigate which atomistic judgments of the experts could predict their holistic judgments on the suitability of a text for STEP. For this, a suitability ratio for each text was calculated based on the suitability judgments of the experts. As mentioned before, the suitability ratio of a text shows the average of "ones" (Suitable) and "zeros" (Unsuitable) given by the experts for each text in the questionnaire. Table 7 summarizes the suitability ratios of the texts in the questionnaire. It can be seen in Table 7 that Text 4 was found unsuitable by a big majority of the experts, with only 4% of the experts finding it suitable. It was followed by Text 7 and Text 8 in unsuitability ranking. On the other hand, Texts 1, 9,

5 and 10 were the texts that were found to be the most suitable for STEP with suitability ratios of .80 or above.

In order to explore what features of these texts might have caused the experts to view the texts suitable or unsuitable; two different analyses were carried out. First, a logistic regression analysis was conducted in order to see which of the judgments on the atomistic features of the texts significantly predicted whether a text would be regarded as suitable or not by the experts. For the analysis, features with a numeric value, from Q3 to Q11, were regressed onto the suitability judgments of the experts. Again Q10 was removed from the analysis due to its high correlation with Q11. The results indicated that the flow of the text (Q11), vocabulary difficulty (Q5), and density of the information (Q7) were the variables that significantly predicted the suitability judgments of the experts χ 2 (4) = 141.40, p <. 001. Nagelkerke's R^2 of .36 indicated a moderate relationship between the prediction and grouping. Prediction success of the model was overall 76.9% (64% for suitable and 86% for unsuitable) (See Appendix E, Tables E1, E2, and E3 for the logistic regression results).

In addition to the logistic regression, a qualitative analysis of the experts' comments for the suitability question (Q13) was carried out to see what observations or judgments of the experts led them to regard a text as (un)suitable. It was observed

Table 7. The Suitability Ratios for Texts

Text Number	Text Code	Suitability Ratio	SD
Text 4	Nov	.04	.20
Text 7	IDis	.15	.35
Text 8	NFoo	.34	.47
Text 2	His	.49	.50
Text 6	ICor	.66	.47
Text 3	Eth	.74	.44
Text 10	IHob	.80	.39
Text 5	Sci	.81	.39
Text 9	NPol	.85	.35
Text 1	Soc	.85	.35

that a very small number of experts had written comments for the texts that they had found suitable. However, the majority of them had written their justifications when they found a text unsuitable. For this reason, in the qualitative analysis, the researcher chose to focus on the justifications that the experts gave for finding a text unsuitable for STEP. Additionally, as there were too few or no comments to analyze for the texts that were found suitable by the majority of the experts, the researcher had to limit the qualitative analyses to texts that were viewed by at least 25% of the experts. For this, the texts with suitability ratios below .75 were chosen. As a result of this, qualitative analyses of the experts' judgments were carried out for six texts, namely Texts 2, 3, 4, 6, 7, and 8. Table 8 summarizes the most frequently given reasons for the unsuitability of these texts. It was observed that the most commonly cited reasons for a text to be regarded as unsuitable were related to their vocabulary and topic related difficulty. This seemed to be the case for Texts 2, 3, 6 and 8.

Text 2 (His) was an extract on "pastoralist way of life" from the History course book at ISU. Those who found it unsuitable justified their choice focusing on vocabulary difficulty and topic specificity. The text presented a lifestyle from 7500 years ago, which is quite unfamiliar to many of the contemporary people. As such, it included topic specific and infrequent vocabulary such as "herders, alluvium, flocks, grazing, transhumance, steppe lands etc." Some of the comments for this text were as follows: "… a dictionary would help them in class, but for an exam No" (Expert 3), "vocabulary is too topic specific." (Expert 19), and "intimidating vocabulary" (Expert 14). Despite judging the text as suitable, one of the experts (Expert 8) raised her concern over the vocabulary difficulty saying "Yes, but with some vocab. changes."

Table 8. The Reasons for Text Unsuitability

Text	Text Code	Nr. of experts to find the text unsuitable	Percentage of experts to find the text unsuitable	Reasons for unsuitability	Nr. of experts to cite the reason
Text 4	Nov	45	96%	Colloquial language Not academic	23 22
Text 7	IDis	39	83%	Illogical cohesion/flow of ideas	31
Text 8	NFoo	29	62%	Culture/topic specific Not academic	15 6
Text 2	His	24	53%	Vocabulary difficulty Topic specificity	18 9
Text 6	ICor	16	34%	Vocabulary difficulty Topic specificity	7 6
Text 3	Eth	12	27%	Topic specificity / abstractness Vocabulary difficulty	9 3

Text 3 (Eth), which was an extract from the Ethics text book at ISU, was a discussion of "making promises from an absolutist point of view". The text included low frequency words such as "presumptive, overridability, construe etc." The use of such low frequency vocabulary and the abstract nature of the subject seemed to be the factors driving the experts to find the text unsuitable. Some comments were as follows: "too abstract" (Expert27), "too philosophical" (Expert 28), and "too much jargon" (Expert 36). Again one expert (Expert 15), despite rating the text suitable, felt the need to express her vocabulary concern saying "Yes, but except the vocabulary."

Text 6 (ICor), an extract from an IELTS practice test issued by Cambridge University Press, was found unsuitable for the same reasons as the previous two texts. It was on different qualities of the "cork tree" and included quite topic specific and infrequent vocabulary such as "bark of a tree, trunk of a tree, buoyant,

sarcophagi, elasticity etc." Some comments were as follows: "the text requires specific knowledge" (Expert 9), "technical terms and too much info." (Expert 38), and "too dense and topic specific vocabulary" (Expert 44).

Text 8 (NFoo), which was a newspaper extract dealing with how footballers can keep fit until the late years of their careers was found unsuitable due to its topic and culture specificity. The text mentioned a number of footballers' names and some specific vocabulary related to the game of football. Some justifications for finding the text unsuitable were: "too culture specific" (Expert 12), "…favors male readers, and background knowledge is required" (Expert 22), and "football is too specific" (Expert 25).

Unlike the previous four texts, which were found to be unsuitable due to their vocabulary difficulty and topic specificity, Text 4 (Nov) was found unsuitable by 96 percent of the experts, for different reasons: being non-academic and including colloquial language. As a narrative extract from the novel "Of Mice and Men", the text includes dialogues with frequent use of informal language. Explaining their justifications for finding the text unsuitable, a big majority of the experts commented on the informal language used and the genre being not relevant to an academic context.

Finally, Text 7 (IDis) was an extract from an IELTS practice test whose flow was intentionally distorted by the researcher to see the responses of the experts. Not surprisingly, most of the experts rejected this text due to its problematic coherence. One expert (Expert 42) wrote "… this question is a tool to measure the seriousness of the respondents to the questionnaire". Another expert (Expert 39) wrote "…serious flaws. Not good for exam unless it is the flaws that the questions are about."

Considering the results of the quantitative and qualitative analyses, it can be said that the suitability judgments of experts were influenced by multiple text features. Both logistic regression and qualitative analyses showed that vocabulary difficulty and the coherence of a text were significant factors to influence the experts' suitability judgments. Additionally, genre, topic specificity, and information density were found to be the other factors to interact with the experts' suitability judgments.

At this point, it might be reasonable to continue the discussion with a comparison of factors affecting overall text difficulty (RQ1a) and text suitability (RQ1b) so that a more complete picture can be visualized. Apparently, some features affected both the difficulty and suitability judgments of the experts. These factors were vocabulary difficulty, coherence, topic specificity, and information density.

As vocabulary was found to be the strongest predictor of text difficulty and the most commonly cited reason for finding a text unsuitable, a deeper look into the interaction between the experts' judgments on vocabulary difficulty, overall difficulty and suitability might be insightful. Looking at the vocabulary difficulty and overall difficulty scores, one can assume that the experts had a tendency to equate vocabulary difficulty with overall text difficulty. The texts that they judged to be lexically easier than most of the texts (e.g. Texts 10, 8, 9, and 1) were also found to be the easiest in terms of overall difficulty (see Appendix C). The texts that were judged to be lexically difficult (E.g. Texts 4, 2, 5, and 6) were also found to be high in overall difficulty (see Appendix C). The only counterexample to this was the incoherent text (Text 7). It can be presumed that incoherence in a text seems to affect the overall difficulty judgments but not necessarily the vocabulary judgments.

When it comes to text suitability, lexically easy texts (e.g. Texts 9 and 10) were found suitable while lexically difficult texts (e.g. Texts 2 and 4) were found unsuitable. The only text that seemed to act differently was Text 5, which was found relatively difficult in terms of vocabulary but still quite suitable for STEP. In short, looking at these results, one can assume that vocabulary difficulty is an important factor influencing both overall difficulty and suitability judgments of the experts.

Another similarity between suitability and overall difficulty judgments was related to text incoherence. Text 7 had a lexical difficulty score of 2.50 and it was categorized as a lexically easy text, as it was below the median score of 3.00 and mean score of 2.97 for lexical difficulty judgments. However, unlike the other lexically easy texts, Text 7 was found to be neither easy (in terms of overall difficulty) nor suitable. In this respect, it would be fair to say that although vocabulary seems to be a main determinant of overall difficulty and suitability of a text, it is not necessarily the only one. A text composed of short and simple sentences with high frequency words might still be judged to be very difficult if it is not coherent. Similarly, a text that resembles the target context in terms of many linguistic factors might still be found unsuitable for a purpose if it lacks coherence.

Information density and topic specificity were also other factors that seemed to influence both the overall difficulty and suitability judgments. However, the finding that judgments on both of these features correlated moderately with vocabulary judgments (see Appendix D) may suggest that it might be the vocabulary judgments driving the judgments on these two other features. The qualitative analyses results also supported this suggestion by further showing that these two factors always seemed to accompany ‘vocabulary difficulty’ as far as the suitability of a text was concerned (see Table 8). Considering these, one may speculate that

experts' information density and topic specificity judgments are influenced by their vocabulary difficulty judgments.

Another point worth discussing is the effect of genre on experts' judgments. The qualitative analysis for RQ1b showed that the suitability judgments were affected by text genre, like in the case of Text 4 (Nov). Experts had mentioned that the texts' colloquial and non-academic language due to its genre were the factors that made it unsuitable for STEP. However, it should be noted that Text 4 was found to be high in overall difficulty as well. A plausible question to ask here is 'Could it be that the genre was the reason why Text 4 was judged to be the most difficult in terms of vocabulary and overall difficulty?' Then, why was it not listed among the factors that affected overall difficulty judgments? The answer to the latter question lies in the nature of the analyses carried out for the two questions. First of all, none of the experts had written any justifications for their overall difficulty judgments on Text 4. Therefore, no qualitative analysis was conducted for RQ1a. Also, the question on text genre (Q2) was categorical in nature and did not enter the multiple regression analysis. Therefore, as genre was not one of the independent variables analyzed for the RQ1a.

To recapitulate, although the analyses suggested the experts' overall difficulty and suitability judgments were affected by multiple factors, vocabulary difficulty and coherence were the predominant factors that were accompanied by topic specificity, genre and information density. Considering the results for the RQ1a and RQ1b, it would be reasonable to say that when it comes to choosing or designing exam texts, test writers need to pay extra attention to the abovementioned text features. Automated textual analysis tools can help test developers in this respect. The following research question (RQ2) was intended to see the correlation between

automated indices and expert judgments to understand which indices can readily replace human judgments in test development.

4.4 The results and discussion for RQ2

This question sought to investigate whether automated textual analysis indices correlated with the average expert judgments (i.e. pooled expert judgments). For this, two separate analyses were conducted. Firstly, for the categorical data from question 1 in the questionnaire (i.e. the purpose of the text), a comparison was made between the expert judgments and two automated tools, namely MAT and Coh-Metrix narrativity percentile scores. Next, for the numerical data, a correlation analysis was conducted to see the correlation between average expert judgments on each textual feature and a number of previously determined indices mentioned in the methodology chapter. Table 9 shows the experts' judgments on the purposes of the texts in the questionnaire as well as Coh-Metrix narrativity percentile scores and MAT analysis results for the texts. It can be seen in the table that the texts that were judged to be informative and descriptive have quite low Coh-Metrix narrativity percentile scores (e.g. Texts 1, 2, 5, 6, and 7). Texts 3 and 4, which are not judged to be mainly informative or descriptive, have high Coh-Metrix narrativity percentile scores meaning that these texts have narrative content. This is also supported by MAT results, according to which Text 3 and Text 4 are story-like texts and thus are labeled as imaginative narrative texts. Additionally, although Text 8 was judged to be an informative and descriptive text by the majority of the experts, 10 of the experts also judged it to be narrative. This is supported by a rather high Coh-Metrix score (48). In addition, MAT classified the text as "general narrative exposition", which is also in line with the judgments of the experts. In other words, this was an

Table 9. The Purposes and Sources of Texts

Text Number	Expert judgments on purpose (nr of experts to say that)	Coh-Metrix narrativity percentile	MAT classification	Source of the text
Text 1 (Soc)	Inform (43) Describe (11)	15	Involved persuasion	Textbook chapter
Text 2 (His)	Inform (34) Describe (25)	16	Learned exposition	Textbook chapter
Text 3 (Eth)	Discuss (36) Compare/Contrast (15)	72	Imaginative narrative	Textbook chapter
Text 4 (Nov)	Narrate (47)	87	Imaginative narrative	Novel
Text 5 (Sci)	Inform (41) Describe (13)	11	Scientific exposition	Textbook chapter
Text 6 (ICor)	Describe (44) Inform (23)	13	General narrative exposition	IELTS preparatory material
Text 7 (IDis)	Inform (32) No answer (7)	6	Learned exposition	IELTS preparatory material
Text 8 (NFoo)	Inform (28) Describe (16) Narrate (10)	48	General narrative exposition	Newspaper article
Text 9 (NPol)	Inform (42) Describe (7)	34	General narrative exposition	Newspaper article
Text 10 (IHob)	Inform (37) Describe (20)	38	Involved persuasion	IELTS preparatory material

informative text using narration to convey the ideas. It is worth pointing out that there are also some differences between the judgments of the experts and the classifications made by MAT. Texts 1 and 10 were judged to be informative/descriptive texts by the experts, and as such one might expect them to be classified as a type of exposition, but they were classified by MAT as “involved persuasion”, meaning that they are persuasive texts written in a way that the writer is quite involved and using the language as if he/she is directly interacting with the readers. This is supported by the existence of such expression as ‘you will soon gain’, ‘all you need’, and ‘of course’ in Text 1 and ‘many of us’, ‘Why do they do

it?', and 'they will look for, say …' in Text 10. Consequently, although they were informative texts, the tones of the writers of these two texts were somewhat direct and the writers sounded as if they were interacting with the readers as in a verbal exchange. The fact that MAT is sensitive to such variations in a text might be regarded as an advantage that the tool offers.

The second analysis for RQ2 involved the comparison of numerical data from expert judgments and automated tool indices through correlation analyses. When checking the correlations between automated indices and average expert judgments, a two-step procedure was followed. First, the correlation between the average expert ratings and the results of the automated tool indices for the 10 texts in the questionnaire was checked. Then, the same procedure was repeated with 7 texts excluding Texts 3 (Eth), 4 (Nov), and 7 (IDis). As mentioned in the previous chapter, the reason to follow such a procedure was that Texts 3 (Eth) and 4 (Nov) were mostly narrative in nature, and Text 7 (IDis) was distorted in terms of coherence. Narrativity in texts was associated with a mismatch between human raters and automated indices (Sheehan et. al, 2010). Additionally, it was thought that the distorted flow of Text 7 was likely to have an effect on other judgments of the experts. Therefore, the researcher expected to see higher correlations between the automated indices and human judgments when 7 texts, rather than 10 texts, were included in the correlation analyses. The following tables from Table 10 to Table 19 show the indices that significantly correlated with the experts' judgments on texts.

Table 10. Correlation between Q3 (Expertise Required to Read the Text) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
Flesch Kincaid GL	.65	.05	CELEX CWF	-.95	.01
CELEX CWF	-.64	.05	BNC Spoken CWF	-.91	.01
			Coh-Metrix L2	-.81	.05
			BNC Written CWF	-.81	.05
			MRCConcreteness	.78	.05

Table 11. Correlation between Q4 (Grammatical Difficulty) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
COCA CWF	- .73	.05	BNC Spoken CWF	-.97	.001
BNC Written CWF	-.64	.05	CELEX CWF	-.96	.001
			BNC Written CWF	-.94	.001
			COCA CWF	-.91	.01
			Coh-Metrix L2	-.89	.01
			Coh-Metrix Concreteness	.83	.05
			Flesch Reading Ease	-.82	.05
			Flesch Kincaid GL	.80	.05
			BNC CWR	-.78	.05

Table 12. Correlation between Q5 (Vocabulary Difficulty) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
No correlations found			CELEX CWF	-.94	.001
			Coh-Metrix L2	-.93	.01
			BNC Spoken CWF	-.93	.01
			MRC Concreteness	.93	.01
			BNC CWR	-.92	.01
			Coh-Metrix Concreteness	.89	.01
			BNC Written CWF	-.89	.01
			COCA CWF	-.82	.01
			Byrsbaert Concreteness	.78	.05

Table 13. Correlation between Q6 (Concreteness) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
Coh-Metrix L2	.86	.01	No correlations found		
COCA AWF	-.83	.01			
Byrsbaert Concreteness	-.81	.01			
BNC AWF	-.81	.01			
TTR CWF	-.76	.05			
MRC Concreteness	-.71	.05			

Table 14. Correlation between Q7 (Information Density) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
CELEX CWF	-.80	.01	CELEX CWF	-.97	.001
BNC CWR	-.64	.05	Coh-Metrix L2	-.95	.001
			BNC Spoken CWF	-.93	.01
			MRC Concreteness	.90	.01
			BNC CWR	-.88	.01
			BNC Written CWF	-.86	.05
			Coh-Metrix Concreteness	.84	.05
			COCA CWF	-.80	.05
			Byrsbaert Concreteness	.77	.05

Table 15. Correlation between Q8 (Topic Specificity) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
BNC CWR	-.70	.05	Nr of words per NP	.85	.05
Nr of words per NP	.65	.05	BNC CWR	-.82	.05
			MRC Concreteness	.79	.05
			BNC Written CWF	-.78	.05
			BNC Spoken CWF	-.76	.05

Table 16. Correlation between Q9 (Culture Specificity) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
No correlations found			No correlations found		

Table 17. Correlation between Q10 (Cohesion) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
No correlations found			No correlations found		

Table 18. Correlation between Q11 (Coherence) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
TTR all words	.64	.05	No correlations found		

Table 19. Correlation between Q12 (Overall Difficulty) Judgments and Indices

Indices correlating when 10 texts entered	r	Sig	Indices correlating when 7 texts entered	r	Sig.
COCA CWF	- .73	.05	BNC Spoken CWF	-.92	.01
BNC CWF	-.69	.05	BNC Written CWF	-.92	.01
			BNC CWR	-.92	.01
			MRC Concreteness	.92	.01
			Coh-Metrix concreteness	.88	.01
			CELEX CWF	-.86	.05
			COCA CWF	-.83	.05
			Coh-Metrix L2 readability	-.83	.05

The first analysis for this question revealed that the judgments of the experts on text purpose (Q1) are in line with Coh-Metrix narrativity percentile scores and MAT classifications. The texts that were found to have an informative purpose by the experts seemed to produce low Coh-Metrix narrativity percentile scores and are mostly categorized as types of exposition by MAT. Considering the commonly accepted fact that university texts are highly informational and non-narrative in

nature (Biber, 1989; Biber et al., 2002; Amjad & Shakir, 2014) and the claim that narrative texts are easier to read than non-narrative ones (Alderson, 2000; Graesser & MCNamara, 2011), one may suggest that these tools can provide valuable data to test developers. Plus, detailed categorization of expository texts by MAT allows one to make more informed choices among different types of exposition. For instance, some universities (e.g. Middle East Technical University) might mainly include core science departments where an abundance of scientific exposition texts are encountered while some other universities (e.g. İstanbul Şehir University) might be composed mainly of social sciences departments and thus may involve more of learned exposition or narrative exposition texts. An analysis of the TLU domain texts through MAT can lead to more relevant decisions in terms of text genres and discourses to include in a test.

The second analysis involved the correlations between the automated tools and expert judgments. For the correlation analyses, the average scores obtained from 47 experts for the texts were used. Strong correlations were found between some indices and expert judgments. It was observed that the automated indices correlated better with the expert judgments when the three experimental cases were taken out. This confirms Sheehan et al.'s (2010) claim that when a text is narrative, the evaluations by automated indices and human raters may not match each other as automated tools tend to underestimate the difficulty of such texts. It is also in line with the findings of Nelson et al.'s (2012) study, which found that automated tools had higher correlations for informational texts than narrative texts. This finding also supports the assumption that cohesion related automated indices used in this study (i.e. Coh-metrix referential cohesion and deep cohesion) cannot yet capture logical flaws that might make a text quite incoherent although these flaws were detected by

the majority of the experts (See Appendix B for the automated textual analysis results of the questionnaire texts).

It was observed that the majority of the judgments had significant correlations with the automated indices when the 7-text sample was used. However, concreteness judgments correlated with some automated indices when the 10-text sample was used but did not correlate with any index when the 7-text version was used. This implies that the incoherence and/or narrativity in texts (e.g. as in Texts 3, 4, and 7) seem to affect the experts' judgments on grammatical difficulty, vocabulary difficulty, information density, and topic specificity but not the judgments on vocabulary concreteness.

Perhaps the most striking finding was that it was the vocabulary related indices that correlated most with the majority of the experts' judgments. For instance, only the vocabulary related indices had significant correlations with the experts' judgments on expertise required to read a text (see Table 10). Apparently, according to the experts, the vocabulary of a text determines how much expertise is required to read it. The negative correlations suggest that the less frequent are the words used, the more expertise on the readers' side is deemed essential by the experts. This confirms Khalifa and Weir (2009) who say that texts written for experts tend to include low frequency words.

It might be regarded as quite perplexing to find in Table 11 that the indices that correlated most with the grammar judgments of the experts were vocabulary related indices (e.g. BNC spoken content word frequency, CELEX word frequency etc.) rather than grammar related indices (e.g. number of words per sentence and syntactic simplicity). The negative correlations between vocabulary indices and the grammar judgments indicate that low frequency and low range words are associated

with high grammatical difficulty. Traditional readability formulas, which take into account factors such as word length or sentence length correlated with expert judgments on grammar, but they did not correlate as strongly as the vocabulary related indices did. The results support Alderson and Kremmel's (2013) and Römer's (2009) previously mentioned claim that grammar and vocabulary are overlapping features. It adds to this claim by further implicating that it is the vocabulary that is the predominant feature in this interaction between grammar and vocabulary judgments. In other words, the experts' judgments on vocabulary difficulty seem to drive their grammar judgments as well. As such, it might be hypothesized that a text with long sentences and complex syntax might still be judged as grammatically easy if it has an abundance of high frequency vocabulary. A comparison of Texts 2 (His) and 8 (NFoot) can illustrate this point. Text 8 was quite similar to Text 2 according to grammar related indices such as sentence length, number of modifiers per noun phrase, syntactic simplicity percentile, and noun phrase density. However, the experts judged it to be much easier in grammar than Text 2. The average grammar judgment scores were 2.09 for Text 8 and 3.23 for Text 2 (see Appendix C). The reason why the experts found Text 2 to be grammatically much more difficult than Text 8 was probably related with the perceived vocabulary difficulty of Text 2. Text 2 had a vocabulary difficulty score of 3.87 while for Text 8 this score was only 2.34. In short, the vocabulary difficulty of Text 2 seems to have caused the experts to regard it as grammatically more difficult than Text 8, although the two texts were very similar in terms of grammar as measured by automated textual analysis indices. Consequently, it can be said that the experts' grammar judgments are largely influenced by their vocabulary judgments; therefore, it is not surprising to observe

grammar judgments correlating with vocabulary related indices rather than grammar related indices.

It can be seen in Table 12 that the vocabulary judgments of the experts correlated with content word frequency and range indices as well as Coh-Metrix L2 and concreteness indices. Not surprisingly, negative correlations were found between the word frequency/range indices and lexical difficulty judgments suggesting that vocabulary of low frequency and range are found to be more difficult. The negative correlation between Coh-Metrix L2 readability and vocabulary judgments means that texts that were found to be lexically difficult are not very readable according to Coh-Metrix L2 redability scores.

As for concreteness, Coh-Metrix L2 readability index was found to have the highest correlation with expert judgments (see Table 13). It was followed by two AWF indices (COCA AWF and BNC AWF), suggesting that the higher frequency of academic words meant less concreteness for the experts. This seems to be in line with Green et al.'s (2010) argument that most of the academic texts deal with abstract subjects. Two concreteness indices (Byrsbaert and MRC) and the TTR for content words also correlated significantly with concreteness judgments. What is interesting here is that the concreteness indices do not have the highest correlations with the concreteness judgments. It remains puzzling why Coh-Metrix L2 readability index and COCA AWF had higher correlations with the concreteness judgments than the concreteness indices did.

The researcher's expectation was for information density judgments to correlate with TTR indices (TTR all words and TTR content words) as the common assumption is that the higher the ratio of types to tokens is, the less repetition of vocabulary there is and thus the more new information is presented in a short text

span. However, the results indicated that word frequency, concreteness, and Coh-Metrix L2 readability were the indices to correlate significantly with experts' information density judgments (see Table 14). Again, it seems that vocabulary judgments of the experts were driving their judgments on information density. It seems hard to explain the reason for the correlation of information density judgments to concreteness and Coh-Metrix L2 readability indices.

There were no tools specifically intended to check for the topic specificity of texts, but some indices correlated significantly with the experts' judgments on topic specificity. Correlating well with the experts' judgments on topic specificity, word frequency and range indices seem promising in this respect (see Table 15). Negative correlations between topic specificity judgments and BNC word frequency and range indices suggest that texts including frequent words from wide range of domains are judged to be less topic-specific. This seems reasonable as the word specificity itself entails moving away from what is common and frequent. This finding is in line with Khalifa and Weir (2009) who state that topic specific vocabulary is usually low in frequency. Other than word frequency, number of words per noun phrase also correlated with topic specificity, although, as a grammar related index, it was expected to correlate with grammar judgments of the experts. According to the results, the texts that had a high average of words per noun phrase were considered more topic-specific by the experts. Lastly, higher concreteness was associated with topic specificity. For instance, Text 2 (on pastoralism) and Text 6 (on cork tree) were quite concrete by the concreteness analysis tools, they were found highly specific by the experts and (see Appendix C). Similarly, Text 9 (on pollution) was found less specific, and it was also not very concrete according to the concreteness indices. However, it would probably be wrong to draw a conclusion like topic specific texts

are always more concrete. Highly specific texts can also be highly abstract as well (e.g. a text on the differences between liberal economic systems and government controlled systems). This interaction between topic specificity and concreteness was probably due to the texts used in the questionnaire.

The results did not indicate any significant correlations between culture specificity judgments and indices (see Table 16). This was not surprising at all as culture specificity is a matter varying across cultures, and as such there are no tools specifically designed to measure this feature. However, it was observed in the qualitative analysis for RQ1b that culture specificity is indeed a factor that influences the experts' text suitability judgments as in the case of Text 8, which was on football. Text 8 was found too culture specific and thus unsuitable for a proficiency exam. As mentioned by Green et al. (2010), cultural specificity might be expected to increase text difficulty and thus reduce suitability. This implies that culture specificity of a text is very likely to influence the performance of test takers and thus needs to be considered when preparing exams. However, in the absence of automated tools to measure texts in terms of their culture specificity, the responsibility falls on the shoulders of test developers to make sure that their exam texts are not loaded with culturally specific content. As Alderson (2000) states, some above-linguistic features of texts, such as culture specificity, can be assessed only by expert human raters, rather than readability formulas or automated tools.

As for the cohesion and coherence questions, the 7-text versions did not yield any correlations between judgments and automated indices (see Table 17 and Table 18). For the 10-text version, TTR all words was the only index that correlated with experts' judgments on flow of ideas in the text (Q11). This finding seems to support Graesser et al.'s (2011) claim that lexical diversity is about the cohesion of a text

because a high TTR means there are a variety of new words in the text that the reader needs to incorporate into the context. The positive correlation between TTR index result and the expert judgments mean that the higher the TTR of a text the more new information and the less repetition there is, which makes the text harder in terms of coherence.

Coh-Metrix deep and referential cohesion indices did not correlate with the cohesion and coherence judgments of the experts. Considering that the majority of experts easily noticed the problems in the flow of Text 7 (IDis) and found the text unsuitable due to its lack of coherence, it can be concluded that these two tools may fail to spot some coherence problems that humans can notice. This seems to be in line with Graesser et al. (2011) who state that current automated tools are not able to fully capture deep metaphors and literary devices in texts. This finding further suggests that automated tools may also fail to comprehend some obvious problems regarding the logical flow of a text. It is worth pointing out that Coh-Metrix referential and deep cohesion indices yielded almost identical scores for the original (coherent) and distorted (incoherent) versions of the Text 7 (IDis) (see appendix F). In other words, the tools did not see any difference between the coherent and incoherent versions of the same text.

The finding that the two cohesion indices did not spot the lack of flow in Text 7 might be taken to imply that test developers need to use these indices cautiously, without too much reliance on their results because these indices seem to take into account the incidence of repetition of words and the number of connectors disregarding the logical connections between sentences. To give an example, a text with sentences like ‘Albert was a very successful student because, first of all, his father was on the bus, and, thirdly, he never did his homework.’ would probably be

rated by referential and deep cohesion indices as somewhat coherent due to the repetition of pronouns and frequent use of conjunctions even though the sentences do not necessarily make much sense together.

As for overall difficulty, again, vocabulary frequency, concreteness indices and CohMetrix L2 readability had the highest correlations with overall difficulty judgments (see Table 19). The negative correlations between the overall difficulty and the vocabulary frequency and range indices suggested that less frequent words are associated with overall textual difficulty. Similary, low Coh-Metrix L2 readability scores were associated with higher difficulty.

To summarize what has been said regarding the correlation between indices and expert judgments, it was seen that the judgments on Q3 (expertise required to read the text), Q4 (grammatical difficulty), Q7 (density of the information), Q8 (topic specificity) and Q12 (overall difficulty) correlated significantly with vocabulary frequency related indices. In other words, vocabulary related automated indices seemed to correlate not only with vocabulary judgments but also with some other judgments of the experts. This finding seems to support the findings for the previous research question. As mentioned in the discussion of the previous research question, the majority of expert judgments significantly correlated with their vocabulary judgments (see appendix D). If vocabulary judgments are driving the other judgments of the experts, then it is no surprise that experts judgments on various features of the texts correlated mostly with vocabulary related indices.

The question of ‘which indices one can use’ does not have a straightforward answer. As seen from the analyses, when texts have a narrative nature or when they are not coherent, indices do not correlate with expert judgments on text features other than concreteness. Additionally, no automated indices correlated with the experts’

judgments on culture specificity of texts. Still, when test writers make sure that the texts they are to use are expository, culturally unbiased and coherent; some indices might provide them with extra information to facilitate the text selection process. In order to determine the extent of narrativity in a text or the type of exposition to choose for an exam text, test writers might make use of Coh-Metrix narrativity percentile and MAT as these two tools seemed to agree with the experts' judgments on text purpose. Additionally, based on the finding that some indices repeatedly correlated with the experts' judgments, test writers might also be recommended to use these indices in text selection and test specification process. The most frequently correlating indices to expert judgments were word frequency and range indices, readability indices and concreteness indices. Of the word frequency indices, the ones that correlated most with the expert judgments were CELEX content word frequency (CWF), BNC CWF for written texts, BNC content word range (CWR) for written texts, BNC CWF for spoken texts, and COCA CWF for written texts. The scores obtained from these indices correlated not only with the vocabulary judgments of the experts but also with the judgments on other features, such as topic specificity, information density and grammar. Of the readability indices, Coh-Metrix L2 readability was the most frequently correlating index with expert judgments (five times). Flesch grade level and Flesch reading ease correlated with expert judgments twice and once respectively, while Lexile did not correlate with any of the expert judgments. The reason why Coh-Metrix L2 readability index correlated better with the expert judges might be due to the fact that a big majority (40 out of 47) of the experts participating in the study was L2 English speakers. The creators of Coh-Metrix L2 readability index claim that this index is better suited for L2 English contexts and traditional readability formulas are intended for L1 speakers. In this

respect, this finding might be considered to be in support of their claim. As for the concreteness indices, MRC concreteness was the most frequently correlating index with the expert judgments. MRC was followed by Coh-Metrix concreteness percentile and Byrsbaert Concreteness. Additionally, COCA academic word frequency (AWF) and BNC AWF indices were found to correlate negatively with the expert's concreteness judgments. That is to say, as the frequencies of academic words in a text increase, the text is perceived to be less concrete by the experts. This implies that the test developers might make use of academic word frequency indices in addition to the abovementioned concreteness indices to better adjust the level of abstractness in their texts.

The main finding that vocabulary frequency and range indices correlating best with the majority of expert judgments seems to support Alderson's (2000, p.73) claim that vocabulary difficulty is commonly accepted to be the most important predictor of text difficulty. As the vocabulary indices seem to be correlating with not only the experts' judgments on vocabulary but also with their judgments on some other features (and thus the overall difficulty of a text), test writers might be advised to make use of the word frequency and range indices mentioned above. However, this is not to say that other indices should completely be abandoned. As mentioned in the literature review chapter, traditional readability formulas based on surface structures like sentence length have repeatedly been found to predict text difficulty. One possible reason why vocabulary judgments stuck out so much in this study was probably that the texts were quite uniform in terms of grammatical difficulty. Taken from university course books, IELTS preparatory materials, and newspaper articles, all the texts in the questionnaire might be taken to be of similar grammatical

difficulty levels in that none was as simple as an intermediate English course book reading text, for instance.

Finally, the unpredictable correlations found between automated indices and the expert judgments might provide valuable information regarding our understanding of language and the nature of expert judgments. For instance, the correlations between (1) concreteness indices and grammar judgments, (2) TTR content words index scores and concreteness judgments, and (3) number of words per noun phrase index scores and topic specificity judgments were not expected as they have not been found to correlate with each other in previous studies.

As mentioned in the literature review chapter, automated textual analysis tools can be used for the document analysis of the TLU domain in order to generate specific and detailed task descriptors (i.e. specifications). The next question deals with a comparison of four corpuses from similar contexts to investigate their similarities to and differences from each other.

4.5 The results and discussion for RQ3

This question aimed to investigate whether corpuses from İstanbul Şehir University (ISU), Boğaziçi University (BOUN), University of Bedfordshire (UB) and IELTS contexts differed from each other with respect to different textual properties. A total of 120 texts (30 texts per corpus) of about 800 words each were analyzed.

First, the corpuses were analyzed through MAT to explore the types of texts in terms of their discourse modes. Then, the corpuses were compared based on their scores on textual analysis indices. For each analysis, outliers were removed as they might lead to faulty decisions (especially Type 1 errors) in statistical significance

tests like ANOVA (Liao, Li, & Brooks, 2016). It was seen that there were never more than four outliers for any corpus in analyses meaning that the each corpus entered the analyses with a minimum of 26 texts.

To explore the similarities and differences of the corpuses with respect to their automated analysis indices scores, a one way between groups analysis of variance (ANOVA) was conducted. It was found that the assumptions were met. For the comparisons, (1) overall readability indices (Coh-Metrix L2 Readability, Flesch Kincaid Reading Ease, Flesch Grade Level, and Lexile), (2) vocabulary frequency and range indices (BNC written frequency and range for content words, BNC spoken frequency and range for content words, BNC written frequency for academic words, COCA written frequency and range for academic words, COCA written frequency for content words, and CELEX word frequency for content words), (3) grammar related indices (sentence length, syntactic simplicity percentile, left embeddedness, noun phrase density and number of words per noun phrase), and (4) concreteness indices (MRC concreteness content words, Byrsbaert concreteness content words, and Coh-Metrix concretness percentile) were used in comparisons. For the purposes of clarity, the results and discussions for each group of indices will be presented separately. In addition to the ANOVA, the corpuses were descriptively analyzed in terms of the different text types they included. For this, text classifications by MAT for each corpus were used.

4.5.1 Corpus comparisons for discourse mode

The corpuses were checked using MAT to see the text classifications. It was found that the all of the texts from the four corpuses were classified as one of four text

types. Table 20 is a summary of the text types for each corpus. It was observed that the most commonly seen text type across corpuses was ‘learned exposition’, which are types of texts that formal and informational expositions focusing on conveying information (Nini, 2014). Examples of such texts are texts from social sciences such as Psychology and Sociology. This was followed by ‘scientific exposition’, texts that are also informational in nature but are more technical than learned exposition. Examples of such texts can usually be found in text books from core sciences such as Physics, Chemistry, and Engineering. ‘General narrative exposition’ was almost equally as common as scientific exposition. General narrative exposition texts are also informative texts but use narration in order to convey information (Nini, 2014). Examples of such texts can usually be found in History course books. The last category was “involved persuasion” including persuasive or argumentative language. Based on the results, it is possible to say that all three types of exposition were commonly seen in the corpuses. Types of texts that were categorized as learned exposition were more common than the other two types of exposition, namely scientific exposition and general narrative exposition. As exposition may involve different types, test developers might be advised to sample from different types of exposition to better represent their contexts. Without using MAT, test developers may risk sampling from only one type of exposition and this, inevitably, would

Table 20. Text type classifications of corpuses by MAT

	ISU	BOUN	UB	IELTS	Total
Learned Exposition	10	12	10	10	42
Scientific Exposition	6	8	11	7	32
General Narrative Exposition	8	6	6	11	31
Involved Persuasion	6	4	3	2	15
Total	30	30	30	30	120

reduce the representativeness of the reading texts in a test in terms of the discourse type.

4.5.2 Corpus comparisons for overall readability indices

Four corpuses were compared to each other in terms of their overall readability scores on four readability indices using one-way ANOVA. Significant differences between the four corpuses were observed in some of the comparisons.

Table 21 is a descriptive summary of the overall readability scores for the corpuses. The ANOVA results indicated that there was a significant difference between the corpuses in terms of their Lexile scores at a $p < .01$ level [$F(3,111) = 3.78$, $p = 0.013$]. Bonferroni post hoc comparisons showed that the mean Lexile score for ISU ($M = 1235$, $SD = 76.98$) was significantly lower than that of UB ($M = 1330$, $SD = 137.46$) (see Appendix G, Table G1). This meant that ISU corpus was significantly more readable (simpler) on average than UB corpus according to Lexile.

The next comparison was made based on the average scores on Flesch-Kincaid grade level. Again a significant difference was observed between the corpuses [$F(3,116) = 4.67$, $p = 0.004$]. The results revealed that average Flesch-

Table 21. Descriptive overall readability scores for corpuses

	Lexile			Flesch Kincaid Grade Level			Flesch Reading Ease			Coh-Metrix L2 Readibility		
	Nr of texts	Mean	SD	Nr of texts	Mean	SD	Nr of texts	Mean	SD	Nr of texts	Mean	SD
ISU	27	1235	77	30	11.56	1.96	30	42.29	9.71	27	12.76	3.57
BOUN	28	1307	103	30	13.02	1.94	30	36.71	11.32	30	12.13	4.84
UB	30	1330	137	30	13.53	2.78	30	39.80	13.33	30	10.98	3.68
IELTS	28	1303	84	30	12.69	1.63	28	46.08	6.20	30	10.33	3.27
Av.		1295	109		12.70	2.21		41.40	10.98		11.52	3.96

Kincaid grade level score for the ISU corpus (M = 11.56, SD = 1.96) was significantly lower than that of UB (M = 13.53, SD = 2.78) (see Appendix G, Table G2). Similar to the previous comparison, ISU was found to be the easiest corpus among the four and was significantly easier than UB.

A comparison of corpuses based on their Flesch Reading Ease scores indicated a significant difference between groups [F (3,114) = 4.38, p = 0.006]. It was found that the average Flesch Reading Ease score for the texts from the BOUN corpus (M = 36.71, SD = 11.32) was significantly lower than that of the texts from the IELTS corpus (M = 46.08, SD = 6.20) (see Appendix G, Table G3). This meant that according to this index, BOUN corpus was significantly more difficult than IELTS corpus.

Finally, the corpuses were compared based on their average scores on Coh-Metrix L2 readability scores. The analysis did not yield any significant differences between corpuses [F (3,113) = 2.29, p = 0.08] (see Appendix G, Table G4). However, it was observed that the average readability scores for the contexts where English was the L2 (BOUN and ISU) seemed to be higher than the contexts where English was the L1 (UB and IELTS) (See Table 21). When one considers that the BOUN and ISU corpuses were compiled from books that are chosen by mostly L2 English speaking instructors (mostly Turkish), and that UB instructors and IELTS test writers are mostly L1 English speakers, it might be expected to see that L2 context corpuses might be more easily readable than L1 context corpuses according to their Coh-Metrix L2 readability scores. To test this assumption, a comparison between an L2 context corpus (ISU and BOUN combined) and an L1 context corpus (UB and IELTS combined) was conducted. The T-test results indicated that L2 context corpus (M = 12.43, SD = 4.26) was significantly easier than the L1 context

corpus (M = 10.65, SD = 3.46) in their Coh-Metrix L2 readability scores [t (115) = 2.48, p = 0.015].

It was observed from the comparisons of corpuses based on different readability indices that these indices produced mixed results in comparing the corpuses with respect to their readability. For Lexile and Flesch grade level, ISU was the corpus that was significantly easier to read than UB. However, according to the results based on Flesch Reading Ease, ISU was not significantly different from other corpuses. This time IELTS was significantly easier than BOUN. Other than this, there were mixed results as to the most difficult corpus as well. According to Lexile and Flesch grade level, UB was the most difficult corpus, while according to Flesh reading ease BOUN was the most difficult corpus. By looking at these results, although there seems to be significant differences found between corpuses on three of the four readability scores, it is hard to tell which corpus is easier and which one is more difficult as the results vary depending on the index used.

Another finding was related to the contradicting results from the Flesch reading ease and Flesch Kincaid grade level indices. Although both Flesch reading ease and Flesch Kincaid grade level are calculations based on the same variables, namely average word length and average sentence length, their results did not match each other in the comparison of corpuses. According to Flesch Kincaid grade level the corpuses were ordered from easiest to the most difficult as ISU, IELTS, BOUN and UB; however, according to Flesch reading ease the ranking from the easiest to the most difficult was IELTS, ISU, UB, and BOUN.

Based on these, one can claim that it is hard to rely completely on these tools. However, based on the results of the analysis that compared a combination of Turkish university corpuses with a combination of UB and IELTS corpus, one can

assume for an L2 English context that Coh-Metrix L2 readability might provide more relevant information. This assumption might be supported by the findings for the RQ2, which indicated that of the four overall readability indices used, Coh-Metrix L2 readability was the most frequently correlating index with the judgments of the experts. The majority of these experts who answered that question were L2 English speakers who were asked to evaluate the texts in the questionnaire in terms of their relevance for L2 English university students. Perhaps this is the reason why ISU and BOUN corpuses had higher Coh-Metrix L2 readability scores than UB and IELTS. That is to say, maybe some of the professors teaching at L2 English universities, namely ISU and BOUN, had tried to choose books that they found more easily readable for L2 English students. Considering this, one can suggest that Coh-Metrix L2 readability can potentially be more useful than other overall readability indices for test writers in L2 English contexts.

4.5.3 Corpus comparisons for vocabulary indices

As for the comparisons of corpuses based on eight vocabulary indices, no differences were found between the corpuses in any of the analyses ($p > .05$) (see Appendix G, Tables G5 to G13). In other words, comparisons of corpuses in terms of their vocabulary index scores indicated that the four corpuses did not differ significantly from each other in terms of the frequencies and ranges of the content words and academic words they had. Table 22 presents a descriptive summary of the vocabulary index scores for the corpuses.

It was observed for RQ1a and RQ1b that vocabulary judgments of the experts seemed to drive their overall textual difficulty and suitability judgments, and for RQ2

Table 22. Descriptive vocabulary frequency and range scores for corpuses

	ISU		BOUN		UB		IELTS	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BNC CWF Written	.82	.26	.73	.18	.82	.14	.80	.19
BNC CWF Spoken	.99	.39	.82	.22	.95	.25	.93	.25
BNC CWR Written	52.83	5.93	52.20	4.78	53.32	4.25	51.56	4.58
BNC CWR Spoken	40.44	6.78	37.81	5.48	40.61	6.66	39.72	5.00
BNC AWF Written	9.31	.90	9.43	1.13	9.35	1.05	9.71	1.28
COCA CWF Written	768	250	735	212	753	157	731	199
COCA AWF Written	9436	922	9546	1231	9631	1312	9764	1270
COCA AWR Written	.33	.06	.33	.04	.33	.05	.32	.05
CELEX CWF	2164	155	2095	125	2086	145	2096	128

that vocabulary related indices had the highest correlations with the expert judgments. The fact that the corpuses did not significantly differ from each other with respect to their scores on these indices might mean these four corpuses do not differ from each other in terms of their lexical demands on the readers. This might be considered good news for the RQ4. The reason for such an assumption is that the relevance of these vocabulary indices in textual analysis was verified by the judgments of 47 experts in RQ2, and now that these corpuses do not significantly differ on these indices, then it seems possible to generate quite specific vocabulary ranges for university level academic English tests.

4.5.4 Corpus comparisons for grammar indices

Comparison of the corpuses on grammar-related indices were conducted using five Coh-Metrix indices, namely average number of words per sentence, noun phrase density, syntactic simplicity percentile, left embeddedness and number of modifiers per noun phrase. Table 23 presents a descriptive summary of the grammar index scores for the corpuses. The ANOVA results did not yield significant differences between corpuses for number of words per sentence, noun phrase density, and number of words per noun phrase ($p > .05$) (see Appendix G, Tables G14 to G16). However, statistically significant differences were found between corpuses for left embeddedness and syntactic simplicity percentile (see Appendix G, Tables G16 and G17). Corpuses differed significantly in their average left embeddedness scores [$F(3,113) = 7.35$, $p = 0.000$]. Bonferroni post hoc comparisons showed that the mean left embeddedness score for UB ($M = 4.54$, $SD = 1.19$) was significantly lower than those of ISU ($M = 5.86$, $SD = 1.29$) and BOUN ($M = 6.08$, $SD = 1.54$). In other words, on average UB texts had fewer words before the main verb than ISU and BOUN texts did. As mentioned before, as the number of words before the main verb

Table 23. Descriptive Coh-Metrix grammar index scores for corpuses

	ISU		BOUN		UB		IELTS	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Syntactic Simplicity Percentile	51.43	18.38	43.33	16.69	36.43	19.18	34.07	14.89
Left Embeddedness	5.86	1.29	6.08	1.54	4.54	1.19	5.46	1.32
Number of Words per Sentence	20.94	3.44	23.26	4.68	23.47	4.88	22.69	2.90
Noun Phrase Density	372.5	18.93	380.9	28.49	370.8	25.68	368.8	24.46
Modifiers per Noun Phrase	.96	.12	.98	.14	1.02	.12	.99	.14

increases, it becomes harder to process the sentences. Based on this, it is possible to say that BOUN and ISU corpuses are significantly more difficult than UB corpus in terms of their Coh-Metrix left embeddedness scores.

Another grammar related index in which corpuses showed significant variance was Coh-Metrix syntactic simplicity percentile [$F(3,116) = 6.06$, $p = 0.01$]. It was found that the mean syntactic simplicity percentile score for ISU ($M = 51.43$, $SD = 18.38$) was significantly higher than those of UB ($M = 36.43$, $SD = 19.18$) and IELTS ($M = 34.07$, $SD = 14.89$). This meant that texts from ISU corpus were significantly easier in terms of syntactic complexity than those from UB and IELTS on average. It is important to note that although significant differences were found between corpuses in terms of syntactic simplicity percentile and left embeddedness scores, both of which are grammatical difficulty indices, the results from the two analyses contradict each other.

Corpuses were compared based on their scores on five grammar indices, and they differed significantly in terms of their average scores on two indices: left embeddedness and syntactic simplicity. According to average left embeddedness scores, UB corpus was found significantly easier than ISU and BOUN. However, according to syntactic simplicity percentile scores, ISU was found significantly easier than UB and IELTS. Although both indices are intended to measure grammatical difficulty of texts, their results are contradictory for the corpuses in this study because looking at the above mentioned results, it is impossible to tell whether, for example, UB is grammatically more difficult than ISU or vice versa. Considering that syntactic simplicity percentile is a more holistic grammar index while left embeddedness is a more minute index focusing on only one feature, one may infer that left embeddeness alone cannot be used to explain overall grammatical difficulty.

A reasonable conclusion to draw would probably be that texts can be grammatically different for different reasons.

4.5.5 Corpus comparisons for concreteness indices

Corpuses were compared based on their concreteness scores derived from three indices: Coh-Metrix concreteness percentile, MRC Concreteness, and Byrsbaert Concreteness. The results revealed that corpuses significantly differed from each other in terms of their average concreteness scores on Byrsbaert Concreteness for content words and MRC concreteness on content words, but not in terms of their Coh-Metrix percentile index scores.

Corpuses differed significantly in their average Byrsbaert concreteness scores at a $p < .001$ level [$F(3,116) = 6.72$, $p = 0.000$]. Bonferroni post hoc comparisons showed that the mean average Byrsbaert concreteness score for IELTS ($M = 2.92$, $SD = 0.21$) was significantly higher than those of ISU ($M = 2.77$, $SD = 0.24$), UB ($M = 52.74$, $SD = 0.17$) and BOUN ($M = 2.71$, $SD = 0.18$) (see Appendix G, Table G19). In other words, on average IELTS corpus was more concrete than all three university corpuses.

A similar result was found for MRC concreteness scores [$F(3,113) = 6.91$, $p = 0.000$]. The analysis indicated that IELTS corpus ($M = 369.73$, $SD = 25.33$) was significantly more concrete than UB ($M = 343.11$, $SD = 15.23$), and BOUN ($M = 347.62$, $SD = 22.39$) (see Appendix G, Table 20).

No significant differences were found between the corpuses in terms of their Coh-Metrix concreteness percentile scores [$F(3,116) = 2.47$, $p = 0.065$] (see

Appendix G, Table G21). Table 24 presents a descriptive summary of the concreteness index scores for the corpuses.

It was found that IELTS corpus was significantly more concrete than the university corpuses according to MRC and Byrsbaert concreteness indices. This is probably due to the fact that IELTS texts are not always academic in nature. As newspaper or magazine articles are more frequently used in IELTS than in university course books, this result was expected. The finding is in line with Green et al.'s (2010) study that found IELTS texts to be more concrete than university course book texts.

4.6 The results and discussion for RQ4

This question sought to investigate the optimal ranges for different textual properties of three corpuses derived from automated indices. The three corpuses were ISU, BOUN and UB, namely a Turkish foundation university, a Turkish State university and an English public university.

90 texts from three corpuses were combined as if it is one single corpus. Then the outliers were removed. In order to generate ranges representing the middle 68 percent of the texts from the corpuses that were analyzed, the Empirical Rule (aka

Table 24. Descriptive concreteness index scores for corpuses

	ISU		BOUN		UB		IELTS	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MRC	354	29.62	347	22.38	343	15.23	369	25.33
Byrsbaert	2.77	.24	2.71	.16	2.74	.17	2.92	.21
Coh-Metrix Concreteness Percentile	38.40	30.81	33.70	24.98	35.50	23.90	51.10	29.18

68-95-99.7 Rule) was used. The range of the values falling within one standard deviation of the mean for each index was taken to determine a representative range.

Similar to the procedure followed in the previous question, optimal range scores for corpuses were derived using overall readability, vocabulary, grammar, and coherence indices. Tables 25 to 28 show the middle 68% ranges for the texts from three university corpuses on these four groups of indices.

Table 25. Middle 68% ranges for the overall readability scores of the university corpuses

Indices	Nr	Mean	SD	Range
Lexile	80	1283	96.71	1187- 1379
Flesh Kincaid Grade Level	89	12.63	2.30	10.33 - 14.93
Flesch Reading Ease	90	39.93	11.73	28.20 - 51.66
Coh-Metrix L2 Readability	85	11.66	3.75	7.91 – 15.41

Table 26. Middle 68% ranges for the vocabulary index scores of the university corpuses

Indices	Nr	Mean	SD	Range
BNC CWF Written	89	.81	.22	.59 - 1.03
BNC CWF Spoken	88	.93	.30	.63 - 1.23
BNC CWR Written	90	52.81	5.50	47.31 - 58.31
BNC CWR Spoken	90	39.81	6.57	33.24 - 46.38
BNC AWF Written	89	9.36	1.10	8.26 - 10.46
COCA CWF Written	88	758	214	544 - 972
COCA AWF Written	90	9504	1200	8304 - 10704
COCA AWR Written	90	.33	.05	.28 - .38
CELEX CWF	90	2115	145	1970 - 2260

Table 27. Middle 68% ranges for the grammar index scores of the university corpuses

Indices	Nr	Mean	SD	Range
Syntactic Simplicity Percentile	90	43.73	18.94	24.79 - 62.67
Left Embeddedness	90	5.54	1.60	3.96 - 7.14
Number of Words per Sentence	90	22.70	4.60	18.10 - 27.30
Noun Phrase Density	89	374	25	349 - 399
Modifiers per Noun Phrase	90	.98	.14	.84 - 1.12

Table 28. Middle 68% ranges for the concreteness index scores of the university corpuses

Indices	Nr	Mean	SD	Range
Byrsbaert concreteness	90	2.74	.19	2.55 - 2.93
MRC concreteness	90	350	25	325 - 375
Coh-Metrix Concreteness Percentile	90	35.84	26.45	9.39 - 62.29

Although the above tables look like a long list of numbers, they might still provide valuable information to the test writers in their text selection and development process. The test writers are recommended not to rely solely on these numbers but rather take them as advisory guidelines in addition to their own judgments. It is important to note that finding a text that fits in every single one of the above ranges might be hard as there are a lot of ranges listed. This situation is similar to what Carl Gustav Jung was trying to express when he said that it is possible to calculate the average weight of a thousand pebbles on a beach, and yet the chances of finding a pebble exactly at that weight is very small. Nonetheless, it is possible to claim that these numbers give us approximate ranges within which most of the texts in three university corpuses fall. Using these ranges, broadening or narrowing them depending on the contextual requirements, or at least focusing on some of the ranges, test writers may develop better adjusted and more detailed

specifications that they can use in their efforts to find or create contextually relevant texts.

4.7 Overall discussion

This study was an attempt to seek ways of increasing the range and accuracy of context validity evidence that could be gathered for test development through the use of human judgments and some automated textual analysis tools. Four questions were asked to analyze the factors that seem to affect the perceived difficulty of a text and its suitability for an EAP test, the level of agreement between human judgments and automated textual analysis tools, the similarities and differences between four corpuses and what ranges of textual characteristics one can come up with by comparing these corpuses.

RQ1a and b: The analyses for the first question (RQ1a and RQ1b) revealed that the human raters' perceptions of difficulty and text suitability are affected by multiple factors, the most dominant of which seems to be the vocabulary of a text. Word recognition seems to be at the heart of the reading activity according to the experts. This finding is in line with the long held view that reading is meaning construction through the recognition of individual words (Perfetti, 2007). It also supports Alderson (2000, p.73) who says 'it has long been known that lexical load is the most significant predictor of text difficulty.' Vocabulary judgments of the experts were found to affect their judgments on other textual characteristics, such as grammar, information density, and topic specificity. Rather than being completely separate characteristics, there seems to be a good deal of overlap between these features, and vocabulary seems to be the most important factor. The suggestion that

vocabulary seems to play a greater role than other features in text difficulty presents a contrast to Alderson's (2000, p.70) statement 'Interaction among lexical, syntactic, discourse and topic variables is such that no one variable can be shown to be paramount'.

As mentioned before, the boundaries between grammatical and lexical knowledge are already questioned (Alderson & Kremmel, 2013; Römer, 2009), and the results of this study seem to support the claim that grammar and vocabulary may in fact be combined into one single category. Individually taken, both vocabulary judgments and grammar judgments correlated well with the overall difficulty judgments, but when put together, vocabulary seemed to explain most of the variance that grammar accounted for, as if grammar were a subcategory of vocabulary in the experts' minds. Therefore, when we say both grammar and vocabulary of a text influences its overall difficulty, we may actually be talking about one single factor.

Topic specificity and information density were also found to correlate with the experts' overall difficulty judgments. However, as seen in the analysis for the RQ2, these judgments of the experts correlated with vocabulary frequency indices. It seems that according to the experts, the use of infrequent words suggests that the text is topic specific and informationally dense. This may be taken to imply that these features may not be addressed separately in textual analyses as they already seem to depend on vocabulary to an important extent. In other words, if the resources are limited and one needs to limit the number of text features to focus on, choosing vocabulary seems to be more promising than choosing any other feature.

However, the expert group chosen for the study might be playing a role leading up to this superior role that vocabulary seems to play in determining text difficulty. All the experts participating in the study had worked or were working at

ISU as faculty instructors and English Preparatory School teachers at the time the data was collected. As a member of the testing office of the ISU English Preparatory School, the researcher had informal conversations with the faculty members and preparatory school teachers about the students, and one common complaint that was voiced was that the students were having trouble understanding the course book texts because of their limited vocabulary knowledge. Apparently, there was an air of discontent about the students' lack of vocabulary knowledge. Therefore, this finding might have resulted from the sensitivity of the participants to vocabulary knowledge of the students. This finding may also be interpreted as suggesting a language learning fact: after a certain level of proficiency, vocabulary might be the predominant factor determining textual complexity as there might be less variance among syntactic features within the materials used for teaching content at universities.

It was also found that cohesion of a text is a quality that experts look for in a readable text. Not surprisingly, experts were able to detect a lack of cohesion in a text and identify it as a factor leading to textual difficulty and unsuitability. The importance of cohesion as a source of text difficulty and suitability supports the well established idea that reading does not only involve the individual recognition of words and understanding individual sentences, but rather putting those meaning units to construct a meaningful whole, forming a macrostructure of the text (Kintsch & van Dijk, 1978). Lack of cohesion in a text naturally makes it harder for the reader to create that macrostructure.

However, as mentioned before in the literature review section, despite providing invaluable data, human judgments are not always reliable as shown by Alderson and Kremmel (2013) and Alderson (1993). Pooling a large number of

experts might reduce the problems that might arise from subjectivity and individual differences; however, this approach is not very practical considering the staffing and time limitations of testing offices in many institutions. That is why the use of automated textual analysis indices has been proposed as a complementary approach to expert judgments. Determining automated tools that can provide useful data in a practical manner was the starting point for the RQ2. For this question, freely available and user friendly tools were chosen. Still, there was the need to limit the number of indices so that the ones that could provide the most useful data would be identified. To check the efficacy of predetermined indices, the texts used in the questionnaire were analyzed by means of the indices, and the results were compared against the judgments of the experts.

RQ2: The first analysis for RQ2 showed that MAT and Coh-Metrix narrativity percentile index seem to agree with the judgments of the experts in terms of the purpose of a text, so it is possible to say that these tools can be used to determine the level of narrativity in a text and what type of exposition the test can be classified as.

It was also found that narrativity in texts reduced the agreement between the indices and the expert judgments. As Sheehan et al. (2010) states, automated indices often underestimate the difficulty of narrative texts vocabulary while overestimating the difficulty of the informative texts. The reason for this is probably that compared with expository texts, most narrative texts include concrete vocabulary and short sentences often in the form of dialogues. This does not necessarily mean that narrative texts are easy to comprehend as the literary language used in such texts often leave a lot to the interpretation of the reader, and many words are used figuratively rather than literally. As Graesser et al. (2011) states, automated tools are

not that advanced to capture these variations in narrative texts and naturally might undermine the difficulty of these texts. This means that test developers should refrain from using these indices for texts that are narrative in nature.

It was also found that the cohesion indices used (Coh-Metrix deep cohesion and referential cohesion) did not correlate with the expert judges. Although Text 7 (IDis) clearly stood out as the most incoherent text in the questionnaire according to expert judgments, the two indices did not rate the text as 'incoherent'. What's more the two indices rated the coherent and incoherent versions of Text 7 almost identically. Although the creators of Coh-Metrix rightly emphasize the significance of cohesion for reading comprehension, their tool fails to capture flaws in the cohesion of a text. The reason for this is likely to be that these two indices check the repetition and semantic relatedness of words (referential cohesion) and the incidence of connectors (deep cohesion) but not the logical connections between sentences that help readers construct meaning. A meaningless text full of connectors and semantically similar words might be found to have high deep and referential cohesion by Coh-Metrix.

Once the narrative and incoherent texts were removed, though, some tools were found to correlate repeatedly with the judgments of the experts. It was mostly the vocabulary frequency and range indices that correlated with the expert judgments. This was not surprising as vocabulary was already found to be the best predictor of overall difficulty (RQ1a) and suitability (RQ1b) judgments. All the judgments other than judgments on cohesion, culture specificity and concreteness, correlated highly with vocabulary frequency and range indices. This probably suggests that although a variety of tools may be developed to analyze different features of texts, experts or human judges seem to have a more holistic approach to

texts, and this approach is primarily driven by their vocabulary judgments. For the experts, if a text includes vocabulary of low frequency and range, it is topic specific and informationally dense in addition to being grammatically difficult and being written for experts. That is the most likely the reason why their judgments on different features of the texts correlated with their vocabulary difficulty judgments and vocabulary frequency and range based indices. As such, one can presumably defend the usefulness of vocabulary frequency and range indices (especially CELEX, BNC and COCA content word frequency and range indices) in that they somewhat reflect the perceptions of expert human judges.

As for the overall readability indices, Coh-Metrix L2 readability seemed to correlate the most with the expert judges, while traditionally oriented formulas such as Flesch Kincaid Grade Level, Flesch Reading Ease correlated only a few times and Lexile did not correlate with the experts at all. This supports Crossley et al. (2011) who found that Coh-Metrix L2 readability was better than other traditional readability formulas at assigning texts to the correct grade level of difficulty. The reason why Coh-Metrix L2 readability correlated with the experts better than the other readability indices might be that all the expert judges were from a Turkish foundation university where the majority of students were L2 English speakers. In other words, the experts might have looked at the texts in terms of their readability for L2 English students. If that were the case, then we could say that Coh-Metrix L2 is efficient in doing what it is designed to do.

None of the grammar indices seemed to correlate with grammar judgments of the experts. As mentioned before, this supports the previously mentioned claim that the experts' grammar judgments were likely to be driven by vocabulary rather than

factors that are traditionally associated with grammar, such as left embeddedness, noun phrase density or sentence length.

Of the concreteness indices MRC and Byrsbaert concreteness for content words correlated with the concreteness judgments of the experts. Additionally, academic word frequency indices (BNC and COCA) also correlated well with judgments on concreteness. As most university texts are rich in academic words, they might be expected to have a higher frequency of academic words than texts from, say, newspapers or magazine articles. That's why, it might be a good idea for test writers to make use of academic word frequency indices, in support of concreteness indices, to better adjust the concreteness level of texts.

RQ3: A comparison of four corpuses (ISU, BOUN, UB, and IELTS) was carried out to see the similarities and differences of these corpuses when analyzed using the automated indices in the previous question. It was seen that all three MAT exposition types were prevalent across university corpuses. The existence of different types of exposition can be regarded as another issue that test developers need to take into consideration. This means that test developers may increase the context validity of their tests by choosing types of exposition that are relevant to their TLU domains.

It was found that the corpuses did not significantly differ from each other in terms of the word frequency and range indices. As mentioned before, vocabulary judgments and indices were found to be highly important factors influencing overall readability of texts according to the expert judges.

The comparison made using overall readability indices produced mixed results. Despite depending on the same variables in the calculations, Flesch Kincaid Grade Level and Flesch Reading Ease produced very different results. According to

Flesch Kincaid Grade Level the most difficult corpus was BOUN while the least difficult was IELTS. However, Flesch Kincaid reading ease tells a different story by showing UB as the most difficult corpus and ISU as the least difficult. It is hard to answer how this might be possible. Consequently, it is hard to identify a clearly more difficult corpus based on the scores derived using overall readability indices.

As for their average concreteness, it was observed that IELTS corpus was clearly more concrete than the university corpuses. As Green et al. (2010) say the concreteness of IELTS texts is probably due to the specificity of topics in university course books, which require a more theoretical and abstract approach to texts than IELTS texts do.

Considering all three universities mentioned accept IELTS as an exemption test, we can say that IELTS seems to have a lot of similarities with the university corpuses in terms of the textual analysis index scores. The only feature on which IELTS was clearly different from university corpuses was in the case of concreteness of texts. As such it is reasonable to presume that a student who gets the required IELTS score by one of these universities might be able to handle most of the linguistic demands of that context. This may also help test writers take IELTS texts as good examples for them in terms of its linguistic features.

RQ4: This question aimed to offer optimal ranges of text characteristics based on the three university corpuses analyzed. As mentioned in the discussion of this question, this study offers many ranges on different textual features, and it is probably very hard and unpractical to find or create texts that fit in all of these ranges. A more practical approach would be eliminating the indices that seem to analyze the same features. Identifying a more practical list of useful indices, test writers can generate text specifications to provide a wide range of context validity

evidence on the linguistic task demands of the texts used in exams. Using the ranges of text characteristics representing three universities, test developers in similar contexts may initially prepare texts that are similar to those in these universities. However, it is always advisable for test developers to analyze their own TLU domains and then develop their context specific test specifications. For the purpose of presenting an example, a table of specifications based on the 90 texts from three universities is given in Table 29. Although there are a high number of ranges for each feature, the indices that correlated most frequently with the experts are chosen for practicality. Taking all of the abovementioned points into account, it can be said that test developers are responsible for ensuring that their exam texts resemble the target context in terms of linguistic demands. Otherwise, their tests would not fully represent the target context adequately and might potentially be a source of what Messick (1995) calls 'construct irrelevant difficulty/easiness'. An exam that does not approximate the target language use context in terms of features that are found to influence test performance is very likely to lack context validity and thus the

Table 29. An example list of specifications on the linguistic task demands of a reading text

Feature	Specifications
Overall Text Purpose	to inform (judged by experts)
Discourse Mode	Exposition (relevant types of exposition using MAT) Non-narrative: a Coh-Metrix narrativity percentile score around 22 (Range: 8-35)
Grammatical Resources	Average words per sentence: 22.7 (Range: 18-27) Syntactic simplicity percentile: 43 (Range: 25-60) Coh-Metrix L2 readability score: 11.7 (Range: 8-15)
Lexical Resources	CELEX word frequency for content words: 2115 (Range: 1970-2260) BNC CWR for written texts: 52.8 (Range: 47-58)
Nature of Information	MRC concreteness: 350 (Range: 325-375) COCA AWF: 9500 (Range: 8300-10700)
Content Knowledge Required	Culturally unbiased and relevant texts on topics that are not too specific (judged by experts)

interpretations based on the results of such an exam would not be satisfactorily justifiable. The responsibility of the test developer is to identify test specifications based on a detailed analysis of the qualities of the TLU domain tasks in terms of linguistic features and then prepare tasks based on those specifications. This study has been able to highlight and provide evidence for the important textual features that should be taken into consideration in test development and the extent to which automated tools can be used in this endeavor. The last chapter following this is a conclusion giving an overall summary of this study.

CHAPTER 5

CONCLUSION

5.1 Introduction

The very nature of assessment involves making inferences regarding the abilities or knowledge of the test takers based on their scores. The justifiability of the score based inferences depends on the amount of validity evidence that test designers can present (Messick, 1996). According to Khalifa and Weir (2009), one type of validity evidence concerns context validity. In the case of TEAP reading texts, it is the test writers' duty to ensure that their reading test tasks are contextually relevant to the target academic domain in terms of linguistic demands such as the grammatical and lexical features of the task, the content knowledge required, the nature of information, the discourse mode, and the relationship between the reader and the writer. If test writers can determine the boundaries of the linguistic demands of the reading texts in the TLU domain, then they can use such information to generate specific task descriptors (i.e. test specifications). Such descriptors can be referred to in the development of reading tests that are consistent and standard in terms of the linguistic task demands placed on test takers. Weir (2005) suggests that expert opinions and document analysis can be used to this end. However, due to the subjectivity involved, expert opinions alone may not suffice as context validity evidence, especially in the case of high stakes testing. As more objective sources, automated textual analysis tools are commonly used in document analysis. However, there are too many tools to choose from and these tools at times contradict with each other. Besides, as Green et al. (2010) states, there are some features (e.g. content knowledge required) that automated tools fail to capture. Therefore, in order to make

the best use of expert judgments and automated textual analysis tools, it seemed reasonable to analyze the nature of expert judgments and to determine the automated tools that can be reliably and practically applicable in support of expert judgments. Plus, a comparison of documents from similar contexts (e.g. university course books) using textual analysis tools would probably shed light into the similarities shared by these contexts and thus might be used to generate some general guidelines representing a larger context.

In order to make the best use of expert judgments in collecting context validity evidence, it was imperative to delve into the process employed by experts when they are to make decisions regarding the difficulty and suitability of reading texts. To this end, expert judges were given a questionnaire which required them to rate the difficulty and suitability of texts along with some other atomistic text features. Atomistic judgments that predicted the overall difficulty and suitability of texts were determined through regression and qualitative analyses. As for the use of automated tools, it was deemed necessary to determine the tools that agreed with the expert judgments so that they could be used with confidence. For this, the texts in the questionnaire were analyzed using predetermined textual analysis tools, and the scores generated by these tools were correlated with the pooled expert judgments (i.e. the average scores generated from the judgments of 47 experts). Next, corpuses from four contexts were analyzed using automated tools, and the scores were compared to each other to see the features in which the corpuses show similarities and differences. Finally, the descriptive statistics from three university corpuses were used to determine the optimal ranges of textual analysis index scores for different linguistic task demands of the reading texts for EAP, and a list of text specifications representing the three university corpuses were suggested.

5.2 Implications

This study offers several implications to those involved in language test development, validation, and research. Firstly, it provides insights into the nature of expert judgments on text suitability and difficulty. It was found that human judgments on text difficulty and suitability are affected by several factors, the most dominant of which seems to be the vocabulary of a text. The vocabulary judgments of the experts also influence their judgments on other text characteristics such as reader-writer relationship, grammatical difficulty, information density and topic specificity. This implies that expert judgments are more holistic evaluations of texts and are primarily driven by vocabulary difficulty. This claim was further supported by the high correlations that vocabulary frequency and range indices yielded with experts' judgments on the features of the texts other than vocabulary difficulty. It may not be so surprising that aspects such as reader-writer relationship, topic specificity and information density are dependent primarily on vocabulary; however, the finding that vocabulary indices, instead of grammar indices, correlated with experts' grammar judgments raises questions regarding the classic segregation of grammar and lexis as two separate components of language. This finding supports the arguments of Alderson and Kremmel (2013) and Römer (2009) who question the existence of any boundaries between grammatical and lexical knowledge.

Secondly, coherence of a text seems to affect the experts' perceptions of text difficulty and suitability. However, although human raters were quick to find an obvious lack of coherence in a text, the two Coh-Metrix cohesion indices used in the study did not perform well in spotting that text as incoherent. These indices check the repetition and semantic relatedness of words (referential cohesion) and the incidence of connectors (deep cohesion) but not the logical connections between sentences that

help readers construct meaning. Therefore, a meaningless or illogical text full of connectors and semantically similar words might be found to have high deep and referential cohesion by Coh-Metrix, as was shown by the present study. Given this, test writers might be advised to rely on their own intuitions than Coh-Metrix referential and deep cohesion indices in the case of determining the coherence of a text.

Thirdly, it was found in the present study that the experts rated the narrative texts to be more difficult than the automated tools did. As such, the present study confirms the claims of Sheehan et al. (2009) who assert that textual analysis tools undermine the difficulty of narrative texts as these tools are not that advanced to capture the covert linguistic variations in narrative texts (e.g. irony, figurative language, jokes etc.). The fact that there are short sentences, dialogues, and high frequency words in a text does not necessarily mean that such a text is easy to read. Given this, test developers might be suggested to use the automated tools to analyze expository texts but to refrain from using them for texts that are narrative in nature.

Fourth, based on the correlations between the tools and human raters, this study provides suggestions for test developers regarding which automated indices they can use more confidently. Based on the findings, one can defend the usefulness of word frequency and range indices employed in this study as they reflect the holistic judgments of the experts to a significant extent. Other than this, of the overall readability indices, Coh-Metrix L2 readability had the highest correlations with the expert judges. Considering that the expert raters were mostly L2 English speakers, we can suggest that in EFL English contexts, Coh-Metrix L2 readability is a more promising tool to check for overall readability than the other traditional readability formulas (i.e. Lexile, Flesch Kincaid Reading Ease and Grade Level), which are

originally designed to take into account L1 English speakers. On the other hand, some tools need to be approached with caution as they did not correlate with the expert judgments. These were the grammatical difficulty indices, the cohesion indices and one index used to analyze information density (i.e. TTR content words).

Next, the comparison of corpuses showed that overall readability formulas might yield mixed results. What is even more interesting is despite depending on exactly the same variables in their calculations (i.e. sentence length and word length), Flesch Kincaid Grade Level and Flesch Reading Ease produced results that contradicted with each other. As such, it seems hard to rely on these two traditional readability formulas. As mentioned above, Coh-Metrix L2 reading ease promises to be a more reliable index as it agreed more with the expert judges.

Finally, by offering specifications representing three different EMI contexts, the study helps test developers in other EMI contexts to be able to compare their own contexts to these three contexts and make relevant inferences regarding to what extent the students in their contexts would be able to handle texts from other similar contexts.

It should be kept in mind that there seems to be limits to human judgments. Although they seem to be quite reliable when it comes to judging the vocabulary difficulty and the coherence of texts, they might not be able to notice differences in syntax unless they are too big to ignore. Plus, the automated tools seem to rely only on the overt features of the texts, but they cannot yet capture the covert features that humans could notice (i.e. culture specificity, logical flaws, variations in narrative texts etc.)

In conclusion, this study advances our understanding of the relations between human textual analysis mechanism and automated textual analysis tools as well as bringing suggestions to the use of both in selecting texts for the assessment of L2 language proficiency at EMI higher education institutions. By using both expert judgments and automated textual analysis tools and by showing how and where they can be meaningfully used, this study suggested guidelines that can be implemented in EMI universities for the development of standard reading tests that represent the TLU domain comprehensively and thus can serve as context validity evidence.

5.3 Limitations and suggestions for future research

This study has some limitations; as such, the results need to be approached with caution. To begin with, the limited number of texts used in the questionnaire might have affected the generalizability of the findings. Only 10 texts were used in the questionnaire due to time and resource limitations. The results would potentially be different had the results from indices and human judges been compared using a larger number of texts. Therefore, it is recommended that, before reaching any hasty conclusions, more research comparing humans to automated tools using a large number of texts varying significantly in many features should be done.

Similarly, the texts did not vary greatly with respect to their syntactic difficulty. Due to the fact that texts were chosen from university course books, IELTS exam samples and newspaper articles, they were quite uniform (B2 or above) in terms of their syntactic difficulty. If there had been more variety and extreme cases in this respect, syntactic difficulty judgments would have been a stronger predictor of overall difficulty and suitability and possibly would have correlated

better with grammatical difficulty indices. This might be the reason why the experts focused more on lexical differences among the texts. It might be taken to imply that vocabulary was the most dominant factor distinguishing the texts used in the study, and therefore, as mentioned above, replication studies with texts showing greater variability in all of the features seem imperative.

Finally, the fact that all the expert judges were instructors from the same institution (ISU) might have caused them to reflect on the questionnaire responses their views or concerns regarding that institution. In other words, the reason why vocabulary stood out as the most dominant factor might be due to a general weakness of the students in that institution in terms of vocabulary, which may stem from various factors ranging from the curriculum or achievement tests to the books used at the English preparatory school of that institution. Thus, it would be more beneficial to use experts from various institutions in studies of expert judgments if the judgments are to be generalized to various contexts.

APPENDIX A

QUESTIONNAIRE GIVEN TO THE EXPERT JUDGES

NAME AND SURNAME: _____

POSITION AT UNIVERSITY: Faculty Instructor SEPP Instructor SEPP Test Writer Other

EXPERIENCE OF WORKING AT UNIVERSITY LEVEL: _____ years.

Dear participant,

We kindly request you to take part in a study that aims at determining the difficulty level of texts to be used in STEP (Şehir English Proficiency Test). We would like you to evaluate the attached texts in terms of their suitability for Freshmen students. In other words, do you think the students in the beginning of the first year should be able to read texts as the ones you are going to read in this set?

You are going to do a language and content analysis of 10 short texts (of about 200 words each). You will read each text and then answer the questions based on the characteristics of the texts. When you are answering the questions, please take into account the language abilities and background knowledge of typical İstanbul Şehir University Freshmen students.

Participation to the survey is optional and your responses to this survey will remain anonymous.

We appreciate your valuable help. Thank you.

TEXT 1.

… Over the past decade or so, research within the social sciences has come to use the Internet more and more (Hewson *et al*., 2003). Three uses will be briefly outlined here. The first can be found in using the Internet to gain relatively straightforward access to data on all manner of worldwide issues. By using a search engine (such as Google) and typing in key words, you will soon gain access to world maps, library catalogues, archived newspapers and journals, official government documents, social movement archives and all manner of cultural phenomena. In many ways this is a good starting point for almost any social research (Gauntlett, 2000), and sometimes it may prove to be all you need: data on the Web is like secondary data that is open to analysis (e.g. crime statistics). A second use can be to deploy research tools on the Internet. The most obvious example here is email interviewing. Having found your sample or special subject, you can ask questions by email and the respondent replies. This can lead to further and fuller questioning. Ultimately, the data gained this way can be handily saved on the computer in carefully designated folders and files (as well, of course, as being printed off). …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 2.

… Pastoralism, which involved the herding of sheep and goats but also cattle, appeared as a way of life around 5500 BCE, essentially at the same time that full-time farmers appeared. The first pastoralists were closely affiliated with agricultural villages whose inhabitants grew grains, especially wheat and barley, which required large parcels of land. Pastoralists produced both meat and dairy products, as well as wool for textiles, and exchanged these products with the agriculturalists for grain, pottery, and other staples. In the fertile crescent surrounding the Mesopotamian alluvium, many extended families farmed and herded at the same time, growing crops on large estates and grazing their herds in the foothills and mountains nearby. These herders moved their livestock seasonally, usually pasturing their flocks in higher lands during summer and in valleys in winter. This movement over short distances is called transhumance and did not require herders to vacate their primary locations, which were generally in the mountain valleys.

A quite different form of pastoralism, often called nomadic pastoralism, also based on the herding of cattle and other livestock, came to flourish in other settings, notably in the steppe lands north of the agricultural zone of southern Eurasia. This way of life was characterized by horse-riding herders of livestock. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 3.

… If we make a promise, for example, we put ourselves in a situation in which the duty to keep promises is a moral consideration. It has presumptive force; and if no conflicting apparent duty is relevant, then the duty to keep our promises automatically becomes an actual duty. What about situations of conflict? For an absolutist, an adequate moral system can never produce moral conflict, nor can a basic moral principle be overridden by another moral principle. But Ross is no absolutist. He allowed for overridability of principles. For example, suppose you have promised your friend that you will help her with her ethics homework at 3:00 P.M. While you are on your way to meet her, you encounter a lost, crying child. There is no one else around to help the little boy, so you help him find his way home. But in doing so you miss your appointment. Have you done the morally right thing? Have you broken your promise? It is possible to construe this situation as constituting a conflict between two moral principles: 1. We ought always to keep our promises. 2. We ought always to help people in need when it is not unreasonably inconvenient to do so. ...

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 4.

… Slim's eyes were level and unwinking. He nodded very slowly. "So what happens?"
George carefully built his line of solitaire cards. "Well, that girl rabbits in an' tells the law she been raped. The guys in Weed start a party out to lynch Lennie. So we sit in a irrigation ditch under water all the rest of that day. Got on'y our heads sticking outa water, an' up under the grass that sticks out from the side of the ditch. An' that night we scrammed outa there."
Slim sat in silence for a moment. "Didn't hurt the girl none, huh?" he asked finally.
"Hell, no. He just scared her. I'd be scared too if he grabbed me. But he never hurt her. He jus' wanted to touch that red dress, like he wants to pet them pups all the time."
"He ain't mean," said Slim. "I can tell a mean guy a mile off."
"'Course he ain't, and he'll do any damn thing I—"
Lennie came in through the door. He wore his blue denim coat over his shoulders like a cape, and he walked hunched way over.
"Hi, Lennie," said George. "How you like the pup now?" Lennie said breathlessly, "He's brown an' white jus' like I wanted."…

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …
(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)
(easy) 1 2 3 4 5 (difficult)
Comments:
_____

5. For typical Freshmen students, the vocabulary of this text is …
(basic/frequent) 1 2 3 4 5 (difficult)
Comments:_____

6. The concepts discussed in this text are …
(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 5.

… In England, by the 17th century, vast quantities of wood had been consumed by the demands of an expanding population and the growth of shipbuilding, construction, and iron manufacture (which required large quantities of charcoal). England's forests were never fully restored, but fuel shortages were alleviated by burning coal in the place of wood. Although there were misgivings about the noxious vapors given off by burning coal, it came to be widely used for domestic heating, and as a source of process heat for the production of beer, sugar, bricks, soap, glass, and iron. More than simply a substitute for wood, by the end of the nineteenth century coal had become the basis of industrial civilization, as the rich coal deposits of Britain significantly contributed to that country's unique position as "the Workshop of the World." Much of the industrial age was the era of coal, as coalfired steam engines powered factories, hauled railroad trains, generated electricity, and propelled ships to distant destinations. Yet, just when coal had established its primacy as the most important energy source for industrial society, hard questions were being asked about the continued viability of coal-based technologies. By the end of the nineteenth century it was becoming evident that stocks of coal, while still large, were being depleted at everincreasing rates. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 6.

… Cork – the thick bark of the cork oak tree (Quercus suber) – is a remarkable material. It is tough, elastic, buoyant, and fire-resistant, and suitable for a wide range of purposes. It has also been used for millennia: the ancient Egyptians sealed their sarcophagi (stone coffins) with cork, while the ancient Greeks and Romans used it for anything from beehives to sandals.

And the cork oak itself is an extraordinary tree. Its bark grows up to 20 cm in thickness, insulating the tree like a coat wrapped around the trunk and branches and keeping the inside at a constant 20 C all year round. Developed most probably as a defence against forest fires, the bark of the cork oak has a particular cellular structure – with about 40 million cells per cubic centimetre – that technology has never succeeded in replicating. The cells are filled with air, which is so buoyant. It also has an elasticity that means you can squash it and watch it spring back to its original size and shape when you release the pressure. Cork oaks grow in a number of Mediterranean countries including Portugal, Spain, Italy, Greece and Morocco. They flourish in warm, sunny climates where there is a 400 milimetres of rain per year, and not more than 800 milimetres. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 7.

Two things distinguish food production from all other productive activities: second, every single needs food for each day and has a right to it; and first, it is hugely dependent on nature. These four unique aspects, one political, the other natural make food production highly vulnerable and similar to all other businesses. However, cultural values are highly fixed in food and agricultural systems worldwide.

Farmers everywhere face major advantages, including extreme weather, long-term climate change, and price volatility in input and product markets. Luckily, smallholder farmers in developing countries must also deal with difficult environments, both natural in terms of soil quality, rainfall, etc., and human in terms of infrastructure, financial systems, markets, knowledge and technology. But, hunger is prevalent among many smallholder farmers in the developing world.

Participants in the online debate argued that our biggest advantage is to address the underlying causes of the agricultural system's ability to ensure sufficient food for nobody. And they identified our dependency on fossil fuels and supportive government policies as the main reasons of this problem.

In order to minimize the risks farmers face, most experts call for greater state intervention. They argue that governments can significantly increase risks for farmers by providing basic services like roads. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 8.

... For AS Roma, the key to shaping its future is not forgetting its past. That past now notably including Francesco Totti, who concluded his remarkable 24-year career with Roma on May 28, in a home match at the Stadio Olimpico against Genoa.

The 40-year-old striker provides an interesting case study of football longevity for the club's director of performance Darcy Norman, who is not your every day sport scientist. Norman is interested in using a supply chain management and systems thinking approach, borrowed from the world of big business and applied to European football. It's an approach based on the idea that knowing that every action sets off a chain of events that will impact performance.

As for Totti, Norman cites a "complex system" that includes "good genetics" and balanced lifestyle that allowed the Roma great to make effective appearances at his age. Totti is very much in tune with his "performance mind set," says Norman. "His ability to read the game, and be at the right place at the right time can compensate for the fact that he may not be as explosive or 1/100th of a second slower," he adds. ... Darcy also notes that there are "definitely things to learn" from Totti, along with the careers of Daniele De Rossi (completing his 14th season at Roma), and Juventus' 39-year-old goalkeeper Gianluigi Buffon. ...

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 9.

… Nations responsible for much of the world's ocean plastic pollution have promised to start cleaning up their act. At a UN oceans summit, delegates from China, Thailand, Indonesia and the Philippines said they would work to keep plastics out of the seas. Some of the promises are not yet formalized and environmentalists say the measures proposed are not nearly urgent enough. But UN officials praised the statement. Meeting in New York, they said it was part of a clear international shift against ocean pollution. Eric Solheim, the UN's environment director, said: "There are quite encouraging signs, with nations taking the ocean much more seriously. Of course, there is a very long way to go because the problems are huge."

It is estimated that 5-13 million tonnes of plastics flow into the world's oceans annually. Much of it is ingested by birds and fish, and fragments of plastic have even been found in organisms at the bottom of the ocean. A recent paper said much of the marine plastic often originates far from the sea – especially in countries which have developed consumer economies faster than their ability to manage waste. The Helmholtz Centre in Leipzig, Germany, estimated that 75% of land-borne marine pollution comes from just 10 rivers, predominantly in Asia. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

TEXT 10.

… Collecting must be one of the most varied of human activities, and it's one that many of us psychologists find fascinating. Many forms of collecting have been dignified with a technical name: an arctophilist collects teddy bears, and philatelist collects postage stamps, and a deltiologist collects postcards. Amassing hundreds or thousands of postcards, chocolate wrappers or whatever, takes time, energy and money that could surely be put to much more productive use. And yet there are millions of collectors around the world. Why do they do it?

There are people who collect because they want to make money – this could be called an instrumental reason for collecting; that is, collecting as a means to an end. They'll look for, say, antiques that they can buy cheaply and expect to be able to sell at a profit. But there may well be a psychological element, too – buying cheap and selling dear can give the collector a sense of triumph. And as selling online is so easy, more and more people are joining in.

Many collectors collect to develop their social life, attending meetings of a group of collectors and exchanging information on items. This is a variant on joining a bridge club or a gym, and similarly brings them into contact with like-minded people. …

1. Which of the following does this text do? (you can choose more than one.)

(1) narrate an event, (2) describe an object, place etc., (3) inform the reader on a point, (4) compare and contrast things or phenomena, (5) analyze a process, (6) discuss a point from different perspectives (7) defend a point

2. This extract is probably taken from … (you can choose more than one.)

(1) newspaper article, (2) magazine article, (3) research article, (4) textbook chapter, (5) book chapter, (6) novel/story, (7) other: _____.

3. This text is written for …

(general audiences) 1 2 3 4 5 (experts)

4. For typical Freshmen students, the grammar of this text is …
(consider passives, compound/complex sentences and phrases etc.)

(easy) 1 2 3 4 5 (difficult)

Comments:

_____

5. For typical Freshmen students, the vocabulary of this text is …

(basic/frequent) 1 2 3 4 5 (difficult)

Comments:_____

6. The concepts discussed in this text are …

(concrete) 1 2 3 4 5 (abstract)

Comments:_____

7. If a text presents a lot of information in a short space, we call it an informationally dense text. The information in this text is …

(not dense) 1 2 3 4 5 (very dense)

Comments:_____

8. The reading of this text requires _____ amount of topic specific knowledge.

(a minimum) 1 2 3 4 5 (a very high)

Comments: _____

9. If a text can be understood by readers from different cultural backgrounds, we call it a culture-free text. The topic of the text is ...

(culture-free) 1 2 3 4 5 (culture-specific)

Comments: _____

10. Sentences in the text are connected to each other …

(very clearly) 1 2 3 4 5 (not clearly)

Comments: _____

11. The flow of the ideas in the text is …

(clear) 1 2 3 4 5 (not clear)

Comments: _____

12. A student at the beginning of the first year will read this text …

(easily) 1 2 3 4 5 (with difficulty)

Comments: _____

13. Do you think this is a suitable text to be used in an exam at the end of the prep. year? Please, circle Yes or No. (Disregard the length)

YES NO

If NO, what is the reason?

APPENDIX B

AUTOMATED TEXTUAL ANALYSIS RESULTS FOR QUESTIONNAIRE TEXTS

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10
Coh-Metrix Readability	11.279	5.562	22.916	12.851	4.942	6.498	10.978	13.937	11.254	11.879
Flesch Reading Ease	54.881	35.906	56.25	97.999	33.943	60.142	29.287	49.879	52.613	59.004
Flesch Grade Level	10.722	14.679	9.016	1.878	16.262	9.280	13.827	12.623	10.331	9.138
Lexile	1150	1450	750	550	1550	1250	1250	1350	1150	1050
MAT	Inv. Pers.	Lea. Exp.	Imag. Nar.	Imag. Nar.	Sci. Exp.	Gen. Nar. Exp.	Lea. Exp.	Gen. Nar. Exp.	Gen. Nar. Exp.	Inv. Pers.
Coh-Metrix Narrativity	15	16	72	87	11	13	7	48	34	38
BNC CWF Written	0.91	0.49	1.1	0.49	0.58	0.79	0.48	0.81	0.88	1.01
BNC CWF Spoken	1.06	0.54	1.58	0.99	0.49	0.82	0.52	1.06	1.11	1.35
BNC CWR Written	56,29	41,31	61,14	50,1	48,27	43,6	51,61	53,74	52,3	56,51
BNC AWF Written	8.31	8.92	5.67	7.71	10.66	9.07	6.78	8.42	9.58	7.2
COCA CWF Written	899	498	992	341	436	758	529	772	809	933
COCA AWF Written	8,39	9,23	5,21	7,66	10,94	8,9	7,09	8,51	9,85	7,42
COCA AWR Written										
CELEX CWF	2.082	1.901	2.392	2.247	1.895	2.008	2.04	2.27	2.219	2.271
Syntactic Simplicity Percentile	68	35	64	72	2	24	69	3	42	55
Left Embeddedness	5.300	6.125	3.071	1.375	11.143	2.75	3.909	6.778	5.417	4.154
Number of Words per Sentence	20.6	25.88	14.5	9.21	31.14	17.75	18.73	25.44	17.75	16.54
Noun Phrase Density	364.078	396.135	339.901	402.715	408.257	370.892	388.350	358.079	408.451	400
Modifiers per Noun Phrase	0.923	1	0.758	0.534	0.908	1.125	1.067	1.167	0.712	0.610
MRC Conreteness	358	418	304	410	412	430	375	359	359	357
Byrsbaert Concreteness	2.63	2.56	2.20	2.92	2.45	2.51	2.67	2.36	2.49	2.64
Coh-Metrix Concreteness Percentile	33	95	2	84	99	90	68	44	36	52
Coh-Metrix Referential Cohesion	4	37	26	15	72	5	7	30	7	3
Coh-Metrix Deep Cohesion	58	47	98	37	98	51	84	35	8	80

APPENDIX C

MEAN SCORES OF THE EXPERT JUDGMENTS ON TEXT FEATURES

	Text 1 (Soc)	Text 2 (His)	Text 3 (Eth)	Text 4 (Nov)	Text 5 (Sci)	Text 6 (ICor)	Text 7 (IDis)	Text 8 (NFoo)	Text 9 (NPol)	Text 10 (IHob)
Q3 (Expertise)	2.74	3.04	2.78	1.74	3.02	2.83	2.42	2.26	2.13	1.79
Q4 (Grammar)	2.36	3.23	2.63	3.19	3.09	2.51	2.33	2.09	2.09	1.89
Q5 (Vocabulary)	2.62	3.87	3.17	3.91	3.40	3.38	2.50	2.34	2.36	2.17
Q6 (Concreteness)	2.06	2.23	3.96	2.43	2.02	1.77	2.72	2.60	1.89	2.32
Q7(Density)	2.80	3.68	2.87	2.22	3.43	3.32	2.84	2.30	2.62	2.19
Q8 (Topic Spec.)	2.02	3.06	2.45	2.51	2.74	3.06	2.49	2.85	2.00	1.79
Q9 (Cult. Spec.)	1.64	2.51	2.21	3.32	2.15	2.00	2.00	2.85	1.70	1.74
Q10 (Cohesion)	1.51	1.87	2.24	3.07	2.06	2.06	4.24	2.72	1.83	1.79
Q11 (Coherence)	1.49	1.81	2.23	2.87	2.04	2.02	4.33	2.72	1.79	1.77
Q12 (Difficulty)	2.28	3.55	3.11	3.83	3.06	3.09	3.16	2.53	2.06	1.98
Q13 (Suitability)	.85	.49	.74	.04	.81	.66	.15	.34	.85	.80

APPENDIX D

CORRELATIONS BETWEEN EXPERT JUDGMENTS ON TEXT FEATURES

		Q3Expertiserequired	Q4Grammar	Q5Vocabulary	Q6Concreteness	Q7Informationdensity	Q8Topicspecificity	Q9Culturespecificity	Q10Sentence connection	Q11Flowofthetext	Q12Overalldifficulty
Q3Expertise required	Correlation	1	,377**	,361**	,270**	,486**	,532**	,201**	,103*	,105*	,356**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000	,027	,024	,000
	N	463	462	463	463	460	462	461	460	463	461
Q4Grammar	Correlation		1	,676**	,278**	,399**	,441**	,375**	,169**	,132**	,598**
	Sig. (2-tailed)			,000	,000	,000	,000	,000	,000	,004	,000
	N		468	468	468	465	467	466	465	468	466
Q5Vocabulary	Correlation			1	,279**	,460**	,491**	,371**	,160**	,139**	,708**
	Sig. (2-tailed)				,000	,000	,000	,000	,001	,003	,000
	N			469	469	466	468	467	466	469	467
Q6Concreteness	Correlation				1	,252**	,316**	,277**	,247**	,264**	,355**
	Sig. (2-tailed)					,000	,000	,000	,000	,000	,000
	N				469	466	468	467	466	469	467
Q7Informationdensity	Correlation					1	,539**	,206**	,107*	,137**	,483**
	Sig. (2-tailed)						,000	,000	,021	,003	,000
	N					466	466	465	464	466	465
Q8Topicspecificity	Correlation						1	,482**	,254**	,239**	,547**
	Sig. (2-tailed)							,000	,000	,000	,000
	N						468	467	465	468	467
Q9Culturespecificity	Correlation							1	,286**	,259**	,406**
	Sig. (2-tailed)								,000	,000	,000
	N							467	464	467	466
Q10Connection between sentences	Correlation								1	,882**	,436**
	Sig. (2-tailed)									,000	,000
	N								466	466	464
Q11Flow of the text	Correlation									1	,447**
	Sig. (2-tailed)										,000
	N									469	467
Q12Overalldifficulty	Correlation										1
	Sig. (2-tailed)										
	N										467

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

APPENDIX E

LOGISTIC REGRESSION TABLES FOR RQ1b

Table E1. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	116,123	1	,000
	Block	116,123	1	,000
	Model	116,123	1	,000
Step 2	Step	11,720	1	,001
	Block	127,843	2	,000
	Model	127,843	2	,000
Step 3	Step	9,220	1	,002
	Block	137,063	3	,000
	Model	137,063	3	,000
Step 4	Step	4,332	1	,037
	Block	141,396	4	,000
	Model	141,396	4	,000

Table E2 Logistic Regression Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	506,129[a]	,224	,301
2	494,410[a]	,244	,328
3	485,190[b]	,259	,348
4	480,857[b]	,266	,357

Table E3 Logistic Regression Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	q11	-,968	,105	84,142	1	,000	,380
	Constant	2,589	,265	95,699	1	,000	13,312
Step 2[b]	q5	-,340	,101	11,340	1	,001	,712
	q11	-,932	,104	80,270	1	,000	,394
	Constant	3,560	,411	75,073	1	,000	35,167
Step 3[c]	q5	-,506	,118	18,554	1	,000	,603
	q7	,355	,120	8,847	1	,003	1,427
	q11	-,970	,106	84,095	1	,000	,379
	Constant	3,144	,430	53,577	1	,000	23,199
Step 4[d]	q5	-,442	,122	13,163	1	,000	,643
	q7	,369	,121	9,241	1	,002	1,446
	q9	-,211	,102	4,296	1	,038	,810
	q11	-,930	,106	76,313	1	,000	,395
	Constant	3,295	,440	56,120	1	,000	26,988
a. Variable(s) entered on step 1: q11.							
b. Variable(s) entered on step 2: q5.							
c. Variable(s) entered on step 3: q7.							
d. Variable(s) entered on step 4: q9.							

APPENDIX F

COH-METRIX COHESION INDEX SCORES FOR THE ORIGINAL AND DISTORTED VERSIONS OF TEXT 7

	Referential Cohesion	Deep Cohesion
Original Version of Text 7	9%	80%
Distorted Version of Text 7	7%	84%

APPENDIX G

CORPUS COMPARISONS WITH ANOVA

Table G1. Corpus Comparisons for Lexile Scores

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	-71,95767	29,82626	,105	-152,0819	8,1665
	UB	-94,81481*	29,33414	,010	-173,6170	-16,0127
	IELTS	-48,14815	29,33414	,621	-126,9503	30,6540
Boun	Şehir	71,95767	29,82626	,105	-8,1665	152,0819
	UB	-22,85714	29,05713	1,000	-100,9152	55,2009
	IELTS	23,80952	29,05713	1,000	-54,2485	101,8676
UB	Şehir	94,81481*	29,33414	,010	16,0127	173,6170
	Boun	22,85714	29,05713	1,000	-55,2009	100,9152
	IELTS	46,66667	28,55175	,630	-30,0337	123,3671
IELTS	Şehir	48,14815	29,33414	,621	-30,6540	126,9503
	Boun	-23,80952	29,05713	1,000	-101,8676	54,2485
	UB	-46,66667	28,55175	,630	-123,3671	30,0337
*. The mean difference is significant at the 0.05 level.						

Table G2. Corpus Comparisons for Flesch Kincaid Grade Level

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	-1,45667	,54760	,054	-2,9266	,0132
	UB	-1,97333*	,54760	,003	-3,4432	-,5034
	IELTS	-1,13000	,54760	,248	-2,5999	,3399
Boun	Şehir	1,45667	,54760	,054	-,0132	2,9266
	UB	-,51667	,54760	1,000	-1,9866	,9532
	IELTS	,32667	,54760	1,000	-1,1432	1,7966
UB	Şehir	1,97333*	,54760	,003	,5034	3,4432
	Boun	,51667	,54760	1,000	-,9532	1,9866
	IELTS	,84333	,54760	,758	-,6266	2,3132
IELTS	Şehir	1,13000	,54760	,248	-,3399	2,5999
	Boun	-,32667	,54760	1,000	-1,7966	1,1432
	UB	-,84333	,54760	,758	-2,3132	,6266
*. The mean difference is significant at the 0.05 level.						

Table G3. Corpus Comparisons for Flesch Reading Ease

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	6,57367	2,71860	,103	-,7260	13,8733
	UB	3,48533	2,71860	1,000	-3,8143	10,7850
	IELTS	-2,79595	2,76672	1,000	-10,2248	4,6329
Boun	Şehir	-6,57367	2,71860	,103	-13,8733	,7260
	UB	-3,08833	2,71860	1,000	-10,3880	4,2113
	IELTS	-9,36962*	2,76672	,006	-16,7985	-1,9408
UB	Şehir	-3,48533	2,71860	1,000	-10,7850	3,8143
	Boun	3,08833	2,71860	1,000	-4,2113	10,3880
	IELTS	-6,28129	2,76672	,150	-13,7101	1,1476
IELTS	Şehir	2,79595	2,76672	1,000	-4,6329	10,2248
	Boun	9,36962*	2,76672	,006	1,9408	16,7985
	UB	6,28129	2,76672	,150	-1,1476	13,7101
*. The mean difference is significant at the 0.05 level.						

Table G4. Corpus Comparisons for Coh-Metrix L2 Readability Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	103,945	3	34,648	2,286	,083
Within Groups	1712,520	113	15,155		
Total	1816,465	116			

Table G5. Corpus Comparisons for COCA CWF (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25718,337	3	8572,779	,200	,896
Within Groups	4837026,740	113	42805,546		
Total	4862745,077	116			

Table G6. Corpus Comparisons for BNC CWF (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,166	3	,055	1,429	,238
Within Groups	4,249	110	,039		
Total	4,414	113			

Table G7. Corpus Comparisons for BNC CWF (Spoken) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,480	3	,160	1,931	,129
Within Groups	9,365	113	,083		
Total	9,845	116			

Table G8. Corpus Comparisons for BNC AWF (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,863	3	,954	,786	,504
Within Groups	137,180	113	1,214		
Total	140,044	116			

Table G9. Corpus Comparisons for BNC CWR (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	49,577	3	16,526	,675	,569
Within Groups	2743,994	112	24,500		
Total	2793,571	115			

Table G10. Corpus Comparisons for BNC CWR (Spoken) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	144,636	3	48,212	1,324	,270
Within Groups	4186,117	115	36,401		
Total	4330,754	118			

Table G11. Corpus Comparisons for BNC AWF (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1699556,640	3	566518,880	,396	,756
Within Groups	164622195,225	115	1431497,350		
Total	166321751,866	118			

Table G12. Corpus Comparisons for COCA AWR (Written) Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,005	3	,002	,627	,599
Within Groups	,286	116	,002		
Total	,291	119			

Table G13. Corpus Comparisons for CELEX CWF Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	119375,633	3	39791,878	2,062	,109
Within Groups	2238334,733	116	19295,989		
Total	2357710,367	119			

Table G14. Corpus Comparisons for Average Number of Words per Sentence

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	116,471	3	38,824	2,358	,075
Within Groups	1893,406	115	16,464		
Total	2009,877	118			

Table G15. Corpus Comparisons for Noun Phrase Density Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2560,997	3	853,666	1,402	,246
Within Groups	70044,114	115	609,079		
Total	72605,111	118			

Table G16. Corpus Comparisons for Modifiers per Noun Phrase

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,068	3	,023	1,212	,309
Within Groups	2,125	114	,019		
Total	2,193	117			

Table G17. Corpus Comparisons for Left Embeddedness

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	-,21384	,34894	1,000	-1,1509	,7232
	UB	1,32562*	,35209	,002	,3801	2,2712
	IELTS	,40133	,34597	1,000	-,5278	1,3304
Boun	Şehir	,21384	,34894	1,000	-,7232	1,1509
	UB	1,53946*	,35501	,000	,5861	2,4928
	IELTS	,61517	,34894	,484	-,3219	1,5522
UB	Şehir	-1,32562*	,35209	,002	-2,2712	-,3801
	Boun	-1,53946*	,35501	,000	-2,4928	-,5861
	IELTS	-,92429	,35209	,059	-1,8698	,0213
IELTS	Şehir	-,40133	,34597	1,000	-1,3304	,5278
	Boun	-,61517	,34894	,484	-1,5522	,3219
	UB	,92429	,35209	,059	-,0213	1,8698

*. The mean difference is significant at the 0.05 level.

Table G18. Corpus Comparisons for Syntactic Simplicity Percentile Scores

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	8,10000	4,48326	,440	-3,9342	20,1342
	UB	15,00000*	4,48326	,007	2,9658	27,0342
	IELTS	17,36667*	4,48326	,001	5,3325	29,4009
Boun	Şehir	-8,10000	4,48326	,440	-20,1342	3,9342
	UB	6,90000	4,48326	,759	-5,1342	18,9342
	IELTS	9,26667	4,48326	,246	-2,7675	21,3009
UB	Şehir	-15,00000*	4,48326	,007	-27,0342	-2,9658
	Boun	-6,90000	4,48326	,759	-18,9342	5,1342
	IELTS	2,36667	4,48326	1,000	-9,6675	14,4009
IELTS	Şehir	-17,36667*	4,48326	,001	-29,4009	-5,3325
	Boun	-9,26667	4,48326	,246	-21,3009	2,7675
	UB	-2,36667	4,48326	1,000	-14,4009	9,6675

*. The mean difference is significant at the 0.05 level.

Table G19. Corpus Comparisons for Byrsbaert Concreteness Scores

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	,06600	,05067	1,000	-,0700	,2020
	UB	,03800	,05067	1,000	-,0980	,1740
	IELTS	-,14300*	,05067	,034	-,2790	-,0070
Boun	Şehir	-,06600	,05067	1,000	-,2020	,0700
	UB	-,02800	,05067	1,000	-,1640	,1080
	IELTS	-,20900*	,05067	,000	-,3450	-,0730
UB	Şehir	-,03800	,05067	1,000	-,1740	,0980
	Boun	,02800	,05067	1,000	-,1080	,1640
	IELTS	-,18100*	,05067	,003	-,3170	-,0450
IELTS	Şehir	,14300*	,05067	,034	,0070	,2790
	Boun	,20900*	,05067	,000	,0730	,3450
	UB	,18100*	,05067	,003	,0450	,3170

*. The mean difference is significant at the 0.05 level.

Table G20. Corpus Comparisons for MRC Concreteness Scores

(I) Source	(J) Source	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Şehir	Boun	7,04700	6,17469	1,000	-9,5351	23,6291
	UB	11,56507	6,34389	,426	-5,4714	28,6016
	IELTS	-15,05500	6,17469	,098	-31,6371	1,5271
Boun	Şehir	-7,04700	6,17469	1,000	-23,6291	9,5351
	UB	4,51807	6,34389	1,000	-12,5184	21,5546
	IELTS	-22,10200*	6,17469	,003	-38,6841	-5,5199
UB	Şehir	-11,56507	6,34389	,426	-28,6016	5,4714
	Boun	-4,51807	6,34389	1,000	-21,5546	12,5184
	IELTS	-26,62007*	6,34389	,000	-43,6566	-9,5836
IELTS	Şehir	15,05500	6,17469	,098	-1,5271	31,6371
	Boun	22,10200*	6,17469	,003	5,5199	38,6841
	UB	26,62007*	6,34389	,000	9,5836	43,6566

*. The mean difference is significant at the 0.05 level.

Table G21. Corpus Comparisons for Coh-Metrix Concreteness Percentile Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5575,425	3	1858,475	2,488	,064
Within Groups	86633,567	116	746,841		
Total	92208,992	119			

REFERENCES

Alderson, J. C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 203-219). Alexandria, VA: TESOL.

Alderson, J.C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535-556.

Amjad, T., & Shakir, A. (2014). Study of Information Generating Linguistic Features in Online University Prospectuses. *Research on Humanities and Social Sciences, 25*(4), 122-127.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bannur, F. M., Abidin, S. A., & Jamil, A. (2015). A Validation Process of ESP Testing Using Weir's Socio Cognitive Framework (2005). *Procedia - Social and Behavioral Sciences, 202*, 199-208.

Benjamin, R. G., (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88.

Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research, 12*(5), 659-67.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics, 27*, 3-43

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly, 36*(1), 9-48.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. V. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329-345.

Byrsbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64*(3), 545-559

Cobb, T. (2003). VocabProfile, The Compleat Lexical Tutor. Retrieved on 5 February 2017 from http://www.lextutor.ca.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84-101

Crossley, S. A., Greenfield, J., & McNamara, D. (2008). New views of validity in language testing. *TESOL Quarterly, 42*(3), 475-493

Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching, 41*(03), 409-429.

D'Este, C. (2012). New views of validity in language testing. *EL.LE, 1*, 61-76.

Dufty, D. F., Graesser, A.C., Lowerse, M. M., & McNamara, D. S. (2006). Assigning grade levels to text books: Is it just readability? Memphis, TN: Institute of Intelligent Systems, Department of Psychology.

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgment. *System, 25*(4), 497-513

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36,* 181-191

Graesser, A. C., McNamara, D. S., Lowerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193-202

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371-398

Graesser, A. C., McNamara, D. S., & Kulikowich, M. J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191-211.

Green, A. (2014). The test of English for academic purposes (TEAP) impact study: Report 1- Preliminary questionnaires to Japanese high school students and teachers. Tokyo: Eiken Foundation of Japan.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-43.

Ilc, G., & Stopar, A. (2014). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing, 32*(4), 443-462.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Knoch, U., & Elder, C. 2013. A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48-66.

Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment,* Malden, MA: Wiley.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Laufer, B., & Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix.* Cambridge: Cambridge University Press.

McNamara, T. (2000). *Language testing*. Oxford, UK: Oxford University Press.

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

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

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

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

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance.* New York, NY: Student Achievement Partners.

Nini, A. (2014). Multidimensional Analysis Tagger 1.2 - Manual. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

O'Sullivan, B., & Weir, C.J. (2011). Test development and validation. *Language testing: Theories and practices*, 13-32. Basingstoke: Palgrave Macmillan.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383.

Plakans, L., & Bilki, Z. (2016). Cohesion features in ESL reading: Comparing beginning, intermediate and advanced textbooks. *Reading in a Foreign Language, 28*, 79-100.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics, 7*(1), 140-162.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99-128.

Sinclair, J.McH. (2004). *Trust the text: Language, corpus and discourse.* London: Routledge.

Sireci, S., & Bond, M.F. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Smith, R. (2000). How the Lexile framework operates. *Popular Measurement, 3*(1), 18-19.

Taylor, L. (2014). A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese University Entrants. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Walt, J. L., & (Jr.), F. S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics, 38*, 191-204.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Williamson, G., Sandvik, T., Stenner, J., & Johnson, A. (2016). Complexity of university texts in the United Kingdom. Retrieved May 15, 2017 from https://metametricsinc.com/research-publications/complexity-university-texts-united-kingdom/?full_article=true