BERKAN ÖZKAN

2021

ITEM ALIGNMENT WITH COURSE OBJECTIVES AND ITEM QUALITY IN AN ENGLISH LANGUAGE PREPARATORY SCHOOL

A MASTER'S THESIS

BY

BERKAN ÖZKAN

CURRICULUM AND INSTRUCTION

İHSAN DOĞRAMACI BİLKENT UNIVERSITY

ANKARA

JULY 2021

Item Alignment with Course Objectives and Item Quality in an English Language Preparatory School

The Graduate School of Education
of
İhsan Doğramacı Bilkent University

by

Berkan Özkan

In Partial Fulfilment of the Requirements for the Degree of

Master of Arts
in
Curriculum and Instruction

Ankara

July 2021

İHSAN DOĞRAMACI BILKENT UNIVERSITY
GRADUATE SCHOOL OF EDUCATION
Item Alignment with Course Objectives and Item Quality in an English Language Preparatory School
Berkan Özkan
June 2021

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction

Asst. Prof. Dr. İlker Kalender (Advisor)

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction.

Asst. Prof. Dr. Armağan Ateşkan (Examining Committee Member)

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction.

Prof. Dr. Halil Giray Berberoğlu, Başkent University (Examining Committee Member)

Approval of the Graduate School of Education

Prof. Dr. Orhan Arıkan (Director)

ABSTRACT

ITEM ALIGNMENT WITH COURSE OBJECTIVES AND ITEM QUALITY IN AN ENGLISH LANGUAGE PREPARATORY SCHOOL

Berkan Özkan

M.A. in Curriculum and Instruction

Advisor: Asst. Prof. Dr. İlker Kalender

July 2021

In this study, the alignment of test items and learning objectives and parameters of assessment items at a private university English language preparatory school were analyzed. For this purpose, all items in quizzes and midterms were examined in terms of their match with learning objectives. In the first stage of the study, item-objective alignment analyses were conducted separately for listening, grammar, speaking and reading domains. In the second stage, item difficulty and discrimination parameters of all items were calculated. Based on these alignment analyses, it was concluded that overall alignment seems quite good; however, few items did not match any objectives and some objectives were tested a lot more than others. Furthermore, there are some items that need revision according to item difficulty and discrimination parameters. At the end of the study, findings were discussed and several suggestions were provided.

Keywords: Item-objective alignment, item difficulty, item discrimination, English language, measurement and evaluation

ÖZET

İNGİLİZCE HAZIRLIK OKULUNDA MADDE-ÖĞRENME HEDEFLERİ

UYUMU VE MADDE KALİTESİ

Berkan Özkan

Eğitim Programları ve Öğretim Yüksek Lisans Programı

Danışman: Dr. Öğr. Üyesi İlker Kalender

Temmuz 2021

Bu çalışmada, özel bir üniversite hazırlık okulunda kullanılan sınav maddeleri ve öğrenme hedefleri arasındaki uyum ve değerlendirme maddelerinin parametreleri analiz edilmiştir. Bu amaç doğrultusunda, ara sınav ve vizelerde kullanılan tüm maddeler, öğrenme hedefleri ile uyum konusunda incelenmiştir. Çalışmanın ilk aşamasında dinleme, dil bilgisi, konuşma ve okuma becerileri için ayrı ayrı madde-öğrenme hedefi uyum analizi yapılmıştır. İkinci aşamada ise tüm maddelerin madde zorluk ve ayırt edicilik parametreleri hesaplanmıştır. Bu uyum analizine bağlı olarak, genel olarak bakıldığında madde-öğrenme hedefi uyumunun oldukça iyi olduğu, fakat az sayıda maddenin öğrenme hedefleri ile uyuşmadığı ve bazı öğrenme hedeflerinin de diğer hedeflere göre çok daha fazla ölçüldüğü sonucuna varılmıştır. Ayrıca, madde zorluk ve ayırt edicilik parametreleri göz önünde bulundurulduğunda bazı maddelerin gözden geçirilmesi gerektiği görülmüştür. Çalışmanın sonunda ise, bulgular tartışılmış ve bazı öneriler sunulmuştur.

Anahtar kelimeler: Madde-hedef uyumu, madde zorluğu, madde ayırt ediciliği, İngilizce, ölçme ve değerlendirme

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Asst. Prof. İlker Kalender for his support throughout the process. This study would not come true without his invaluable feedback and guidance.

Thanks are extended to all the Academicians from whom I took courses during my study at İhsan Doğramacı Bilkent University. They all contributed to me a lot.

I am also grateful to my Director Dr. Taner Yapar for his support throughout the process and for providing me with the assessment tools and other materials to be used in my thesis.

I owe special thanks to my wife Ezgi Özkan for her endless support throughout this study and I am deeply grateful to my little daughter Derin who has always been my motivation. Finally, I also want to thank my parents who have always supported me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

Teaching and learning a foreign language is a process and they both have various components. Not only students and teachers but also the curriculum, clear objectives, measurement and assessment are vital parts of language teaching and learning. The unity of these components is of utmost importance so as to achieve high level of target language.

Having clear objectives is one of the major components of language teaching. Without clear objectives, curriculum designers might have difficulty in planning a well-established curriculum. Having clear objectives is also crucial because teachers need to know their aims before they start teaching so that they can plan their lessons accordingly. McKeown (2001) claimed that objectives are the first step to assessment. She also said that in order to measure if and what students learn, first the teacher needs to clarify what the students are expected to learn. And this can be achieved with learning objectives. If one of these components is not taken into consideration, it might interfere with teaching and might cause obstacles for teachers and learners. Lawson (2016) stated that the initial stage of designing a course is to prepare the learning objectives. Objectives help teachers stay focused on the intended outcomes of the course. Curriculum is also based on objectives as teachers have to teach what is intended.

Another essential component of teaching a language is measurement and assessment. Assessment is one of the aspects of teaching which needs to be consistent with the curriculum and objectives. In other words, assessment tools should test what is taught and show to what extent learning objectives are covered. There are basically two types of assessment which are formative and summative assessment. According to Scriven (1967), formative assessment is conducted to follow whether the learning process is going on as expected and summative assessment is conducted to understand whether the learning takes place effectively at the end. Brown (2003) described formative assessment as a type of assessment conducted during the teaching and learning process and summative assessment as an assessment which is conducted at the end of a process to measure what students have learnt. Similarly, McMillan (2000) explained the term formative assessment as a type of assessment which is conducted while students are learning the knowledge and summative assessment as another type conducted at the end of a learning process.

Assessment is applied to understand whether students meet the intended objectives or not. Moreover, assessment is also used to check the effectiveness of curriculum. Learners' performance shed light on the curriculum and based on the assessment of learners, necessary changes and improvements may be done in the curriculum. Popham (2001) exemplifies these views by saying that assessment is conducted to learn what the students know, and if students master the curricular goals, the teacher can continue teaching with the following skills. If some problems about teaching are identified thanks to assessment, modifications can be done on the curriculum. It is important to have a link between objectives and assessment materials; otherwise, to what extent learning objectives are met cannot be known for sure.

Thus, in the light of the statements above, the main purpose of this study was to determine the alignment level between learning objectives and test items. Furthermore, item quality of tests through item analysis was also conducted.

Background

Language Proficiency

If we consider English as the target language, in the simplest terms, language proficiency is having enough knowledge of English for a certain goal. If people successfully complete a task which requires English, they can be considered as proficient for that task. However, language proficiency is also defined from different perspectives as researchers and experts have not reached a consensus on its definition. Chiswick and Miller (1995) defined language proficiency as the ability of speaking, or in a more general term fluency, as most of the data available in the target language are on speaking rather than other skills. McCauley and Christiansen (2019) described language proficiency as an ability to transform the information into something meaningful. Graham (1987) focused on the proficiency exams and scores stating that English language proficiency is measured by some commercial tests and test takers' proficiency level is based on their exam performance; therefore, proficiency may be defined differently in some commercial tests. This means that proficiency exams have different scoring systems and even the same scores may indicate different proficiency levels depending on the exam. Thus, although they have different definitions, measurement and assessment validity is a concern for all parties.

English is widely used in all areas such as international business, computer science, information technologies and more which justifies the claim that it has become the Lingua Franca (Randall, 2015). Being able to communicate and use English effectively is one of the most important qualifications that a person should have. In the same way, international companies require high levels of English language proficiency.

In countries such as Turkey, where English is not the native language, English medium of instruction is highly used in different levels of education including higher education. In order to be successful in preparatory class at any English-medium university, students have to reach the required English proficiency level. Thus, not only teaching but also assessing English has become necessary.

At university level, language proficiency can be usually verified by an exemption certificate or an exam conducted by universities' prep schools. However, those who do not possess an exemption certificate have to study English at preparatory level provided that the university is an English-medium institution. At the end of preparatory class, students either take the university's language proficiency exam or an international one. While some universities measure the language proficiency of students themselves, other universities use international English proficiency exams such as Test of English as a Foreign Language (TOEFL) and The International English Language Testing System (IELTS). Although international exams have well-documented evidence for reliability and validity, exams prepared by preparatory schools may lack such evidence.

Learning Objectives

A learning objective is the piece of information on a certain topic or subject which should be learned by the learners (Alonso et al., 2008). Towns (2010) posited that as soon as the aims are decided and set, developers can write learning objectives. He added that objectives are the expected goals of an educational activity which should be measurable. In this definition of learning objectives, being measurable is important since each objective should be measured and assessed during and at the end of the program.

Independently of the topic, effective teaching has a lot of components one of which is learning objectives. In their study Combs et al. (2008) suggested that planning a course has different stages and the initial stage is writing the learning objectives, so prior to teaching, learning objectives should be clearly defined to determine teachers' goal in a specific course. While writing the learning objectives, teachers and other professionals involved in this process have the chance to determine the most crucial parts of the course. They added that writing learning objectives is not only fundamental but it is also difficult; thus, some kinds of training may be provided for teachers to help them write better. From this point of view, we can infer that the curriculum and materials should be developed accordingly and they should be consistent. Writing clear objectives is fairly important during the planning of an English course as it is in other fields.

Domains in Language Teaching

While learning a language, it is important to master all the components and domains of a language. Generally, there is a misconception that grammar is the main focus of language and knowing grammar means being good at that target language. Higgs (1985) opposed to this idea stating that both grammar and vocabulary are essential for effective communication, but being only good at grammar or vocabulary is not enough. However, in order to master a language, all skills, which are listening, reading, writing and speaking, should be adequate for communication. In order to meet the standards of globalization, communicative competences have become essential in language learning. Learning can be ameliorated thanks to importance given to communicative competences (Prapaisit de Segovia & Hardison, 2008). Without doubt, grammar and vocabulary are indispensable parts of a language as well. Having good knowledge of grammar and wide range of vocabulary also fosters other skills.

Hinkel (2006) suggested that mastering only one skill is not useful for language learners as communication requires various skills such as listening, speaking and comprehension at the same time. Therefore, one should focus on all skills rather than just one. From Hinkel's point of view, we can conclude that language is mainly for communication and in order for effective communication all skills have to be improved.

Although four skills, which are listening, speaking, reading and writing, are considered as a whole in language learning, each skill might be measured separately or some of these skills can be measured at the same time. In order to ascertain a person's language proficiency of a specific skill, it might be better to test the candidate in more than one skill since fours skills of English are closely connected and using more than one skill while measuring can provide more reliability (Powers, 2010).

Students' mastery in language skills is also linked with other components of language. While writing learning objectives in foreign language education, the level of students is of utmost importance. Teachers or curriculum developers have to make sure that learning objectives are appropriate for students' language level. As English is a widely-used language throughout the world, there are some prominent frameworks in English language teaching for standardization purposes and Common European Framework of Reference (CEFR), Cambridge English Teaching Framework and Global Scale of English (GSE) are some of them. Among these frameworks, Common European Framework of Reference (Council of Europe, n.d.), which plays a vital role in English teaching, is highly accepted. CEFR offers a descriptive chart, six different proficiency levels which are A1, A2, B1, B2, C1 and C2, and descriptors for each one. Besides, it provides some suggestions for curriculum and reflection. North (2014) stated that one of the main purposes of CEFR is to make teachers and language experts ponder on present applications with the aim of promoting easier communication.

Item Quality

Not only teaching but also measurement and assessment are very crucial in foreign language teaching and learning. Institutions, schools, teachers and trainers should measure students constantly to make sure that students make progress and also program is effective. Otherwise, they cannot be sure about the efficiency of the program and cannot ascertain the strengths and weaknesses of the curriculum either. Black and William (1998) also supported constant measurement by stating that formative assessment has more powerful effects on learning. Taking the assessment results into consideration, teachers can make the necessary changes on curriculum and the program.

Although applied in many educational areas, assessment in foreign language education is not the same as other areas. It has its own characteristics and methods which are basically linked with language acquisition and psychometrics (Alderson et al., 2017). As language is a very wide field of study, it has various assessment alternatives. Brown and Hudson (1998) stated that language assessment differs from other fields of study in terms of intricacy of language and variety of exams. They added that since 1950s, some alternative assessment methods have arisen and although there have been discussions on the reliability and validity of these assessment tools, they all have their unique pros and cons.

Assessment has two major components which are reliability and validity. In order for an assessment tool to be effective, it should be reliable and valid so that we can get accurate results. According to Fulcher and Davidson (2007) validity of an assessment tool means measuring what is aimed to measure. In other words, if institutions, schools and teachers prepare exams with the aim of measuring specific skills and the exam items measure what is intended, the assessment tool is considered as valid. Henning (1987) referred to reliability in terms of consistency in measurement. He defined reliability as getting the same scores after measuring the students' skills multiple times.

Developing test items is a challenging task and testers need to pay attention to a lot of aspects. A good exam item has some qualities. According to Boland et al. (2010), the main purpose of good test items is to make a distinction between the students who learn the subject well and who do not. Before writing test items, testers need to pay regard to what needs to be measured.

Item analysis has a lot of goals, but the most important one is its benefits for making some changes or improvements on assessment tools. Item analysis enables teachers to determine the quality of items so that they can do the necessary changes on exam items or the exams as well. Based on the results of this analysis, teachers may discard some items which are not appropriate for assessment or make some changes on some items to make them better for assessment. Better test items mean better assessment and it paves the way for better teaching.

Item parameters are of high importance in terms of determining the item quality. Item difficulty and item discrimination analyses are used frequently to determine the item quality. Gronlund (1977) emphasized the use of item quality by saying that the main aim of tests is not only to grade students but also rank them and in order to achieve this ranking, variety is needed in test scores. Thus, item difficulty plays a vital role to achieve this. Ebel and Frisbie (1991) refers to the importance of item discrimination indices for reliability of a test. They claimed that using items with better discrimination values make the test more reliable. They also suggested that item difficulty and discrimination are also related to each other and items with moderate difficulty provide the best discrimination indices in a test. Popham (2003) also supported this view by stating that items with moderate difficulty work best to create a variety in test scores which contributes to item discrimination.

Problem

Assessment can be one of the aspects affecting students' motivation and learning. Ames (1992) supported this view by stating that evaluation methods are among the most significant factors that have an impact on student motivation. Thus, type of evaluation may affect learning positively or negatively. If students cannot establish a link between the exams and what they learn, it might negatively affect the learning process.

On the other hand, teachers also have a crucial role in teaching especially in motivating the students towards learning which can be fostered by their assessment methods as well. Johnston and Goettsch (2000) claimed that language teachers need to know more than the rules of that language. Mastery of language taught is a must; however, language teachers should know about student motivation and assessment as well. Thus, assessment is one of the indispensable aspects of language teaching along with the curriculum and teachers' skills and competence.

Educators may use variety of ways to assess students' learning and progress. There are some alternative assessment methods such as portfolios, presentations or project assignments. However, testing students by using standardized tests is a widely used method that defines students’ language skills according to predefined standards. In testing process, learning objectives play a vital role as well.

The alignment of test items and learning objectives is highly crucial for valid assessment. As D’Agostino et al. (2008) stated the agreement between the items of assessment tools and learning objectives is quite important to confirm the validity. Institutions and teachers should define learning objectives quite clearly before they design the curriculum and start teaching process and measure students’ intended skills accordingly. If they do not correspond to each other and students are not tested based on what they learn, they might think that they do not have to learn what is taught in order to be successful in their exams.

On the other hand, if they correspond to each other, it enhances the quality of language teaching as it is considered as proof that teachers are on the right path. Before students move on to the next language level, not only teachers but also students can be sure that they successfully complete the level and ready for the next level. This correspondence also provides feedback about students’ learning and the program or curriculum. Towns (2010) stated that universities may benefit from assessment to enhance the programs and courses.

The best way to be sure of content validity is this alignment between the test items and learning objectives; however, in some cases, test items may not match with the course objectives but still it does not mean that testing process is completely not effective. Sometimes, test items may not match with course objectives, but instead they may match with the objectives of the course books used in classes. So, while checking the alignment between the objectives and test items, course book objectives might also be taken into consideration in case there are some items which do not match with course objectives; that’s why, course book objectives were also included in this study.

The quality of assessment items is also highly essential in assessing the students’ level. Exam items lacking quality is a major problem for institutions and conducting exams with high validity and reliability is quite demanding (Xing & Hambleton, 2004). Teachers might measure whether the learning objectives are attained or not; nevertheless, if the quality of questions is not good enough, measurement and assessment cannot be done properly. This leads to misevaluation of students and also prevents teachers from assessing their and students’ performance. Scott et al. (2006) refered to the importance of exam items’ quality in their study. They posited that the quality of exams about measuring the students’ mastery in a specific topic is measured by two components which are reliability and validity. Therefore, in order to get a clear idea of students’ language levels and effectiveness of the program, the quality of assessment tools should also be taken into consideration.

In the light of the statements mentioned above, the inconsistency between the learning objectives and assessment tools in teaching English can be considered as a problem which leads to ineffective measurement and assessment, misevaluation of students’ skills and failure to make the necessary changes on curriculum. Therefore, there is a need to evaluate regularly the correspondence of items that test English skills to the learning objectives.

Purpose

The main purpose of this study was to determine the alignment level between the learning objectives and formative and summative assessment tools of English preparatory class in a higher education institution in Turkey. At this stage, to what extent the assessment tools correspond to the learning objectives was analyzed. Another purpose of this study was to analyze item quality. To this end, item analysis was conducted to examine item indices such as difficulty and discrimination.

Research Questions

The study addresses the following questions:

1. What is the item alignment level between learning objectives and assessment tools in a private university preparatory school?
    a. What is the item alignment level between listening domain learning objectives and assessment tools?
    b. What is the item alignment level between grammar domain learning objectives and assessment tools?
    c. What is the item alignment level between speaking domain learning objectives and assessment tools?
    d. What is the item alignment level between reading domain learning objectives and assessment tools?
2. What is the item quality of English language assessment tools in a private university Prep School?
    a. What are the item parameters of listening subtests?
    b. What are the item parameters of grammar subtests?
    c. What are the item parameters of reading subtests?

Significance

As mentioned in the previous sections, being proficient in English is a requirement for most of the universities, especially for English-medium universities. Students in such universities should have proficiency in English to be able to follow the courses in their degree programs. For those who study at preparatory class, universities try to provide good education to prepare the students not only for their academic but also professional life. For this purpose, every aspect of language education has to be well-organized and systematic.

By analyzing the consistency between learning objectives and assessment tools in a private university in Turkey and item quality, this study might be of great help to both teachers and institutions. To start with, this study may raise awareness on this issue. All institutions and teachers probably conduct measurement and assessment procedures during or after their educational programs. However, not all of them may take the consistency between the objectives and assessment into consideration.

If assessment tools measure the intended objectives correctly, it will provide an opportunity for institutions to update their programs, curriculum or assessment tools based on the results. Therefore, not only this specific institution but also others can accurately define the areas where students show weaker performance based on the assessment results and they can modify the program accordingly. Jeong (2013) asserted that accurate assessment provides teachers with the opportunity to explain and use the results.

This study might help the institution in which the study was conducted to analyze the exam items in various aspects in detail. The institution may benefit from the results to have more clear understanding of its assessment tools. Based on the detailed analyses, necessary changes might be done and revised assessment tools may provide better assessment. In the same way, this study might raise awareness for other institutions in terms of item analysis of the assessment tools as well. Item analysis can be considered as an indispensable aspect of assessment as teachers should make sure they use the items which are of good quality. Wainer (1989) suggested that item analysis is essential for test developers to produce assessment tools of high quality.

With the right assessment tools and items of good quality, teachers can measure the students' performance accurately. This is essential for the institution where the study was conducted as it teaches foreign language and the preparatory class of this institution consists of different levels and students' levels are determined based on the students' performance in assessment procedures. According to Purpura (2016), exams conducted during or after the courses are one of the ways to collect data about the students in language assessment. He also stated that after collecting data thanks to exams, teachers may interpret these data and use them as evidence to make some decisions. This is one of the most crucial aspects of teaching as teachers usually build on previous knowledge and determine the flow of the course accordingly, and if they cannot evaluate students accurately, they may skip some of the important issues or topics which may hinder learning. Teachers should make sure that students progress gradually learning all the topics thoroughly.

Another possible significance of this study is that with accurate assessment, students' levels can be determined more accurately which enables better teaching as the level of students in the same classroom will be the same. As Sawyer (1996) mentioned in his article, placing students correctly is highly crucial for the efficiency of a course. The value of correct placement can be understood by looking at the benefits and also the losses of incorrect placement. To illustrate, if students are not evaluated based on the objectives, they may fail and repeat the same level although they do not need to. Another case might be that if they are not tested on what they learn, they may pass to next level thanks to their prior knowledge which makes the assessment unreliable and invalid.

CHAPTER 2: REVIEW OF RELATED LITERATURE

Introduction

The purpose of this study was to determine whether there is an alignment between the learning objectives of courses taught and assessment tools of English preparatory class in a higher education institution. Another purpose of this study was to analyze the quality of assessment tools through statistical item analysis.

The first part of the literature review focused on different approaches to language assessment, measurement instruments of language skills, developing test items and analysis of items. In the second part, different views on objectives-test item correspondence and the results of studies investigating this correspondence were presented. At the end of this Chapter, a short summary was given.

Language Assessment

As McNamara (2014) stated, language exams are quite important for people and they benefit from these exams in various fields. They need exam scores to receive better education, to be employed or even to immigrate to countries where the target language is spoken as a mother language. Similar to this view, Brown (2012) suggested that language assessment is necessary for people and it can be used for different goals. Based on their language exam scores, people may get promotion, find a job and even get a citizenship. Thus, language assessment is often used while important career decisions are made for people.

As language assessment is quite crucial, institutions, schools and teachers should know the basics of assessment so as to use it for the good of students. They should also know how to collect reliable data from students via assessment and use the data to improve student achievement (Stiggins, 2014).

Approaches to Language Assessment

Language assessment has different approaches. As a traditional way, institutions apply teacher-based assessment during which teachers evaluate students' language skills. Besides teacher-based assessment, computer-based assessment has become popular and both have some benefits and drawbacks for teachers and students.

Davison and Leung (2009) proposed some benefits of teacher-based assessment. They asserted that teachers play an active role from the beginning to the end of the process and through this process, they collect variety of student artifacts. They added that this assessment type also provides teachers with the opportunity to give immediate feedback. Ishihara (2009) mentioned another advantage of teacher-based assessment by stating that it allows teachers to understand students' skills and knowledge which enables them to prepare their course and syllabus accordingly.

Besides these advantages, there are also some concerns about teacher-based assessment. Teachers are the only authority while measuring students' language skills in the classroom. While conducting quantitative assessment, the analysis indicates that students' results are consistent with standardized English tests. Nevertheless, while conducting qualitative assessment, teachers' grades may not be reliable as there are many factors affecting students' performance such as their attitudes, characteristics and teachers' way of grading. These might create inconsistencies during the assessment process (Llosa, 2011).

Computer-based language assessments have also been developed recently to measure students' performance. However, there have been discrepancies on the usefulness of computer-based language assessment. Computer-based language assessment integrates technology to improve and ease the assessment process (Winke

& Isbell, 2017). Similar to this view, Ma (2013) posited that thanks to computer-based assessment, language learning is not restricted to location and time and it focuses more on individualized learning. In a large-sized classroom, giving immediate feedback can be difficult for teachers. Graff (2003) also stated that tests are available on demand at any time and computer-based assessment provides immediate feedback.

On the other hand, Perrin and Mayhew (2000) raised some concerns regarding the computer-based assessment by stating that students might cheat, print the tests and share them with others.

Measurement Instruments for English Proficiency

Language skills such as reading, writing, listening and speaking can be measured by various exams types and they have effects on the success of students in language assessment (Palmer, 1991). Depending on the purpose, teachers may conduct different types of exams to assess students' language skills and it is their responsibility to decide on the best assessment tool depending on the teaching context (Coombe et al., 2007). Haladyna (1997) also suggested that teachers' main goal is to measure to what extent the objectives of teaching have been attained and in order to achieve this, they ought to find the most appropriate tools and item formats that match with their purpose. In some cases, more than one format can fit teachers' goals.

Language has four basic skills which are listening, reading, speaking and writing. Besides these skills, grammar is also crucial in language teaching. In order to measure these skills, various exam types and alternatives can be used.

Reading is one of the most important language skills as we read variety of pieces of writing in our daily lives. According to Coombe et al. (2007), although reading is an essential language skill, it is hard to measure reading skills of students as they can only be observed indirectly thanks to sub-skills. Brown (2003) argued that in order to have good reading skills, students need to have basic bottom-up strategies to deal with words or phrases and top-down strategies to understand the texts. Moreover, they also need background and cultural knowledge for deeper understanding and analysis.

Assessment of grammatical accuracy is usually done by integrating grammar into four skills. However, there are also some basic item types which are sentence unscrambling, error correction, fill-in-the blanks and sentence completion. Besides these traditional methods, there are some new ways of grammar assessment such as redefining the construct and partial scoring (Larsen-Freeman, 2009).

However, Nozadze (2013) opposed some of these assessment types by mentioning that people never fill in the gaps in their daily lives or they do not need multiple choice. Instead they communicate with each other via productive skills which are writing or speaking. Thus, he believed that these skills should be assessed rather than using multiple choice and fill in the gaps item types as they are much more important than other skills while communicating with people. On the other hand, grammar is also believed to be important to create meaning and although it is not a productive skill, it is still necessary for meaningful communication and people need grammar accuracy as well.

Recently, there has been a shift towards more authentic contexts in the assessment of listening skill which means that more real-life situations and tasks are being used (Taylor & Geranpayeh, 2011). Similarly, in order to provide more real-life environment for students, videos or pictures as well as audios have come into use and even in TOEFL-IBT test they have started to use a photograph of the speaker or other visual aids during the listening tasks (Lynch, 2011).

Listening assessment is quite similar to that of reading and the only difference is that students listen to a text rather than read it. Multiple choice format is commonly used in listening assessment and besides this, a performance task during which students describe a picture is also used (Mead & Rubin, 1985). Weir (1990) also mentioned multiple choice format in listening assessment; however, he claimed that this format is not valid anymore based on the feedback collected from teachers and testers.

While assessing students' writing skills, essays have a vital role in this assessment process. Marzano (2006) stated that rather than short answer questions, essays are more practical and functional in assessing students writing abilities as they demand more complicated writing skills.

When compared to other skills, speaking assessment is more different than other skills as there is lots of interaction. Moreover, while assessing speaking skills, it might be more difficult to maintain reliability and validity due to the human factor and even the dialogues can be totally different when the same teacher assesses the students (Luoma, 2004). Similarly, Seong (2017) stated that since speaking is a complex skill to measure, which components should be included in assessment is still a matter of discussion. However, using analytical rubrics instead of holistic is widely accepted regardless of the class size.

While assessing speaking, most common methods are conversations and interviews with students which are quite traditional. According to Brown (2003) conversations and interviews besides other speaking assessment methods are highly correlated with listening as students both have to understand the target language and speak. He also referred to micro and macro skills in speaking assessment and suggested using these skills as a checklist during the assessment. Bygate (2009) also claimed that testing of speaking ability is related to performance and oral proficiency. Fan and Yan (2020) suggested rather than using traditional methods to assess speaking skills, new methods which benefit from technology should also be used.

Developing Test Items

Multiple-choice items are frequently used in measurement, and there are some fundamental issues regarding multiple-choice item writing. Kehoe (1995) posited that there are three main aspects of multiple-choice items that need to be addressed carefully which are the question stem, options and continuous item development based on the results of item analysis. On the other hand, Frary (1995) focused on what should be avoided while writing multiple-choice items and suggested that typing errors, grammar mistakes and contradictory distracters should be avoided in order to have multiple-choice items of good quality.

Despite their frequent use, multiple-choice items are also criticized because of some aspects. Although there is consensus on the need to write multiple-choice items which measure students' comprehension and application, most of the multiple-choice items measure students' basic knowledge and memorization skills (Aiken, 1982). In her study, Scouller (1998) supported this view by saying that even students perceive multiple-choice exams as an assessment method for lower levels of cognitive skills. However, despite the fact that there are some subjective aspects of multiple-choice items such as determining the distracters, they are marker friendly and easier to evaluate as markers are not influenced by any prejudice (Weir, 1990).

Another type of frequently used assessment is short answer questions. According to McDaniel et al. (2007), short answer questions are more effective than multiple choice questions and they can also be useful to boost learning besides evaluation. Likewise, McDermott et al. (2014) posited that short answer questions are more likely to increase later test performance when compared to multiple choice questions although they require more effort to conduct and evaluate. Brown and Hudson (1998) suggested both pros and cons of short answer questions. They claimed that short answer questions are not time consuming to check and grade; however, as teachers evaluate only a couple of words, it does not show the whole picture.

Brassil and Couch (2019) mentioned the negative aspect of true-false questions by saying that students' chance of giving the correct answer is one in two while this ratio is one in four in multiple choice questions; thus, student success can be explained by chance to a certain extent. Brown and Hudson (1998) also emphasized the high chance factor and added that this type of questions must be well-written for discrimination among the students.

Students' writing skills are mostly measured by essay or paragraph writing and paragraph writing can be considered as the first step towards essay writing. That's why mostly low level students are tested with paragraph writing and high levels are tested with essay writing. Gronlund (1977) stated that essay questions are the ones that give students a chance to reflect their freedom and creativity and he mentioned two different types one of which is restricted response questions to which students are expected to give limited answers. Another type is extended response essay questions and students have no limit while answering these questions. Brown

(2003) also mentioned four different types of writing assessment which are imitative, controlled, responsive and extensive writing.

Speaking assessment might be considered difficult because when compared to other skills like listening or reading, it is more subjective. Luoma (2004) supported this view by saying that in speaking exams students' performances might not be the same and ratings might change due to the human factor and referred to interrater reliability which means students are assessed by different evaluators and given similar scores. Brown (2003) suggested some ways to assess speaking skills some of which are picture-cued tasks, direct response tasks, translation and question and answer tasks.

Analysis of Items

Item Analysis

After conducting an exam, it is of high importance to analyze the exam items based on the students' answers and this provides data on how effective the exam items are. This data can be obtained via item analysis (Gronlund, 1977). According to Moses (2017), item analysis can be defined as obtaining some indices which can be used to summarize, assess or compare items, especially by using item difficulty and discrimination indices. Another explanation of item analysis by Valette (1977) is that item analysis results are obtained by analyzing the students' responses on tests. Nitko (2004) expressed similar opinions by stating that item analysis is a process in which testers gather data from students' responses and use the data to reach some conclusions and item analysis presents statistics including item difficulty and discrimination indices.

There are different views on item difficulty and discrimination values and thus, the indices change. To illustrate, while Ebel and Frisbie (1991) have three categories for item difficulty, Hopkins and Antes (1990) have five different categories, so the values and labels differ. Furthermore, Magno and Ouano (2010) also have three different categories for item difficulty, but the values are different when compared to Ebel and Frisbie (1991). For example, they categorized items as easy with the value .80 or higher, on the other hand, Magno and Ouano (2010) consider items as easy with the value .76 or higher.

If Classical Test Theory is used, item analysis can only be done after the exams are conducted as student responses are needed for analysis. In order to assess the item quality, expert opinion can be considered to have an idea on the items before conducting the exam. Furthermore, in terms of item quality, also distractor analysis can be conducted after the exam.

Item Difficulty

Item difficulty is an important factor in exam preparation. Item difficulty is the proportion of test takers who answer an item correctly to the whole test takers (Miller et al., 2009). Item difficulty is affected by different variables and question stem and distracters are among these. The similarity between the correct answer and distracters may affect the difficulty level of an exam item (Ascalon et al., 2007).

Bachman (1990) asserted that in order to evaluate the reliability of a test, assessment of item difficulty is quite crucial and it is conducted by many institutions. While preparing a test, it is crucial to prepare it in accordance with students’ levels. Testers should balance the item difficulty levels for the validity of exams (Sung et al., 2015). Similarly, Henning (1987) stated that difficulty of test items should be arranged based on the test takers’ abilities and a test should not be either too easy or difficult.

Scheuneman and Gerritz (1990) approached the issue from a different point of view and stated that item difficulty is not only about the quality of test items but it is mostly about the students' performance in the exam. Similar questions may have different difficulty levels and sometimes it may be hard for testers to identify the reasons.

Gronlund (1977) focused on the benefits of item difficulty analysis by stating that thanks to item difficulty analysis, teachers may identify the most difficult items in the exams. Then, rather than spending time with easy items, teachers may focus on the most difficult ones for feedback after the exams.

Item Discrimination

Gronlund (1977) explained how to calculate the item discrimination index. Based on Classical Test Theory, it can be easily calculated by subtracting the number of students who answered correctly in the lower group from those in the higher group and dividing by the number in each group. Item discrimination is a measurement of the relationship between item score and the overall test score and if an item fails to discriminate low achievers and high achievers in a test, it affects the reliability negatively and causes measurement error (Wu et al., 2016). Item discrimination index is used in order to show the discriminating power of an item. Colman (2008) described item discrimination index as the figure showing the difference in responses of lower group students and higher group students. Gronlund (1977) also explained item discrimination values and stated that the highest value for item discrimination is 1.00 and negative values can also be obtained when more low achievers than high achievers give correct answer for an item. He added that such items should not be used in the exam anymore.

Item discrimination has a significant role in assessing item quality as items in exams aim at providing information about students' ability which exams tend to measure (Miller et al., 2009). For a good item discrimination value, distractors are also considered important and McDonald (2007) supported this view by claiming that distractors have a very crucial role for a good item discrimination value and determining and measuring the students' abilities are mostly based on distractors.

Objectives-Test Items Correspondence

The correspondence between learning objectives and test items is significant and this view has been supported by a number of scholars and their studies. Sireci (1998) stated four different elements of content validity which are domain definition, domain representation, domain relevance and appropriateness of test construction procedures. D'Agustino et al. (2008) proposed that in order for a test to be valid, the correspondence between the objectives and items is an inevitable condition. Without this correspondence, it is highly probable that the assessment process will not be successful. Cox and Graham (1966) emphasized the importance of content in measurement. They claimed that students are supposed to be tested based on the content of the course, but different testing methods can be applied for each skill that students need to develop.

Another view supporting this correlation is that the test items prepared to measure the intended learning outcomes should be consistent with teachers' goals. Otherwise, assessment process will probably fail. To illustrate, if the objective is improving speaking skills of students, a grammar-based test will probably not be useful and effective to assess this skill (Carroll, 1980).

On the other hand, Sugianto (2016) refers to the uniqueness of test items and objectives. What is meant by uniqueness is that each institution has its own curriculum, objectives and test items. The items might match the objectives and curriculum of an institution, but this does not mean that same items will match the objectives of another institution. The correspondence between objectives and test items is highly related to curriculum and syllabus of a course.

Another importance of objective-test item correspondence is that if learning objectives are well-defined and if test items cover the whole objectives, the performance of students in the test can accurately represent students learning. On the contrary, when the objectives and test items do not perfectly correspond, the assessment is not valuable for students' learning (Hambleton & Eignor, 1979). Similarly, La Marca et al. (2000) stated that assessment tools should let the students show the information they learned and skills they gained based on the learning objectives and only then can their performance be evaluated.

In a study conducted by Alimi and Ellece (2003) on item-objective correspondence at a university, they aimed at analyzing the correspondence between learning objectives and exams; however, they observed that teachers even did not have course objectives and exams were carried out based on what was taught in the class. They argued that this situation could result in students' learning to some extent, but made it hard for students to apply them in real-life situations. In another study, Hartzell (1984) also mentioned the importance of objective-test item correspondence. He conducted a study on the curriculum and test correspondence in two different school districts and suggested that with the help of correspondence level of objectives and test items, it is possible to come to a conclusion on the effectiveness of not only the program but also the school.

There are different alternatives for the analysis of item-objective alignment. D'Agostino et al. (2008) stated that among the alignment analysis methods, rating and matching techniques are the most common and traditional ones and most other techniques, although they are different in many aspects, also depend on these two. Rovinelli and Hambleton (1977) created a rating model for the alignment of items and objectives which had a Likert type scale (three-point) and subject-matter experts rated the degree of alignment using “1”, “0” or “-1”. Hambleton (1984) suggested another type of alignment analysis which is matching. In this method, learning objectives are matched with the best item representing it. Another method was proposed by Porter (2002), and in this method, there is a classification table and different SMEs do the item-objective matching individually without affecting each other. At times, this matching process might require some discussion among the experts as some of the item-objective matchings can be controversial and open to discussion (Webb, 2007).

As the literature suggests, the importance of correspondence between objectives and test items is undeniable. This correspondence should be paid attention for better assessment and teaching process.

Summary of Literature Review

Not only is learning a foreign language necessary to get good grades at school but it is a must for every field of people’s lives. As it is an important process, all parts of language teaching and learning should be addressed for better results.

As most of the literature suggests, teaching starts with learning objectives. This is the first step towards effective teaching. Learning objectives are important as they make teachers be aware of their goals and what they are teaching. Thanks to objectives, not only teachers know what and why they are teaching but also students know what they are learning.

However, writing clear learning objectives is not the only aspect of effective teaching. Without the correspondence between the objectives and test items, validity of tests will be affected as the content will not match. If learning objectives are not tested, effective teaching cannot be achieved or teachers cannot be sure of students' learning. Moreover, if clear objectives are thoroughly tested in an exam, this exam can be considered as the true indicator of student learning.

To put it in a nutshell, in order to get good results in language teaching, besides teaching methodologies and approaches, learning objectives and testing of these objectives are crucial whichever testing method is used. These issues made me choose my topic for my dissertation and in Chapter 4, I first matched the items with objectives to analyze the agreement level and then analyzed the item parameters.

CHAPTER 3: METHOD

Introduction

The purpose of this study was to determine whether there is an alignment between the learning objectives of courses taught and assessment tools of English preparatory class in a higher education institution. Another purpose of this study was to analyze the quality of assessment tools in terms of item difficulty and item discrimination. In this chapter, methodology of the thesis was presented. First, research design was explained. Then information about the school context and participants was provided. Furthermore, instrumentation, methods of data collection and data analysis were clearly defined.

Research Design

For the first part of this study, which is determining the agreement level between learning objectives and test items, content analysis was conducted. For the second part of this study, item difficulty and discrimination analyses were conducted to analyze the quality of test items.

In this study, the content of exams was analyzed in detail to provide data about objective-item alignment and quality of items. According to Neuendorf (2016), content analysis has become an empirical method which is still used in various areas by many scholars. Krippendorff (2018) also supported this view by stating that the roots of content analysis traced back to past; however, the current form of content analysis differs from the past in terms of aim and method by especially being more empirical. Mayring (2004) defined content analysis as the standardized analysis or assessment of communication. Another definition provided by Weber (1990) is that content analysis is the classification of material by dividing it into smaller pieces of information to make it more manageable. In this study, thanks to content analysis, the agreement level of exam contents and learning objectives was determined and analyzed in detail.

For the second part of this study, descriptive analyses of item parameters were conducted. The purpose of the study was to reach a conclusion only about the institution in which this study was carried out. The purpose was not to learn about other institutions, that’s why inferential statistics was not used. Thus, there is no intention to generalize the results of this study to other contexts.

Context

This study was conducted at a private university preparatory school in Ankara, Turkey. There are six faculties in the university where the data came from. The university accepts students based on their university entrance exam scores and all students who enroll in any bachelor’s program except Turkish Language and Literature and Visual Communication Design have to attend the preparatory school unless they have an exemption certificate. Students can certify their English proficiency by submitting a valid score from exams such as TOEFL IBT or TOEFL ITP (Institutional Testing Program). Otherwise, they have to attend one-year (the duration might change based on the students’ performance) preparatory program before continuing their undergraduate studies.

The preparatory school has three different language levels which are A, B and C levels. Each level also has different classes for students who fail and repeat the level. These levels are determined based on the preliminary exam which is conducted at the beginning of fall semester. Students are placed into different levels on the basis of their scores. All levels follow the same type of program (10 hours of Main Course, 10 hours of Reading & Writing, 5 hours of Listening & Speaking classes) with different curriculum and materials. The detailed information about the levels is as follows:

A Level

A level has two sub-levels which are A Foundation and A Beginner. If the students get 49 or below from the preliminary qualifying exam, they are placed in A foundation level. If the students get a score between 50 and 64 from the preliminary qualifying exam, they are placed in A beginner level. If students get 399 or below from TOEFL ITP exam, they are also placed in A beginner level. Both A foundation and A beginner level students are required to have an average of 70 to become eligible to pass the next level (B Pre-intermediate). If not, the students have to repeat the level (AR) for one more semester.

B Level

If the students get 65 and above from the preliminary qualifying exam, they take the TOEFL ITP exam as the second step. If they get 500 or more, they can take the departmental courses. If their grade is between 400 and 449, they are put in B level. These students are required to have an average grade of 65 to become eligible to pass to the next level (C intermediate). If not, they have to repeat the level (BR) for one more semester. If they have an average grade of 80 or more, they can take the TOEFL ITP exam at the end of the semester as well.

C Level

If the students get 65 and above from the preliminary qualifying exam, they take the TOEFL ITP as the second step. If they get 500 or more, they can take the departmental courses. If their grade is between 450 and 499, they are put in C level. These students are required to have an average grade of 65 to become eligible to take the TOEFL ITP test. If not, they have to repeat the level (CR) for one more term.

For all these different levels, both quizzes and midterms are conducted each year to assess students' learning. In each level, students have to take four quizzes and three midterms. Besides these exams, portfolio tasks are also given for alternative assessment. Each exam has its own weighting and the sum of all exams comprises the overall grade of students at the end of the semester.

Participants

The participants of this study were B level students of 2016-2017 academic year spring semester preparatory level of a private university in Ankara. In this study, purposive sampling was used and all the students in B level participated in this study. B level students were chosen intentionally as most of the students study in B level in the preparatory program. Before B level, students study in A level and A level students cannot take the TOEFL ITP exam. There are also some students starting from B level. So, majority of students study in this level before they take the proficiency exam. If they do not succeed in the proficiency exam at the end of B level, they progress to C level. The most crowded semester of B level is spring semester, so the study was conducted using the spring semester data. The sample of this study comprise of 458 students in total. Of the participants, 226 (49%) were males and 232 (51%) were females. Moreover, 452 students were in their first and 6 students were in their second year of English preparatory class. The great majority of preparatory school students were 18 years old. As the study was conducted at a private university, some students were on scholarship while others paid the whole tuition fees. Sixty-three (14%) students were on full scholarship, 40 (8%) of them were on 75% and 201 (44%) of them were on 50% scholarship. Furthermore, 154 (34%) students did not have scholarship.

Instrumentation

In order to respond to first purpose of the study, learning objectives of each course were analyzed and the alignment level of objectives and test items was determined. Each course has a syllabus including the course content, course objectives and materials. The syllabi were prepared by the Curriculum Development Unit of the university. Measurement and Evaluation Unit prepares the exams based on the course content and learning objectives.

The university applies basically two different exam types which are quizzes and midterms. They vary in the number of questions, duration and content. The exams used to collect data are prepared by Measurement and Evaluation Unit members of the preparatory school in the university. Once the items are written and the exam is ready, unit members review the exams and necessary changes and improvements are made. Finally, each exam is proofread by a native-speaker to prevent any language mistakes. As the last step of exam preparation, the exams are reviewed by Curriculum Unit members to make sure that they are aligned with the curriculum and course content. The detailed information about exams can be found in Table 1 below.

Table 1

Preparatory School Exam Types Per Semester

Exam Type	Number of Items	Content	Duration
Quiz 1	40	Listening, reading, vocabulary, grammar	50min
Quiz 2	40	Vocabulary, grammar, writing	50min
Quiz 3	20	Listening, reading, vocabulary, grammar	50min
Quiz 4	-	Speaking	15min
Midterm 1	80	Listening, reading, vocabulary, grammar, writing	130min
Midterm 2	80	Listening, reading, vocabulary, grammar, writing	130min
Midterm 3	100	Listening, reading, vocabulary, grammar, writing	150min

The total weighting of abovementioned exams is 73 out of 100 (each quiz 7 out of 100, midterm 1 10 out of 100, midterm 2 15 out of 100 and midterm 3 20 out of 100). Other assessment tools used in preparatory school are portfolios (7 out of 100), assignment and projects (10 out of 100) and reader exams (10 out of 100).

Most of the test items in the exams stated above were multiple-choice items. The other items were true-false questions, matching, paragraph or essay writing or speaking questions. The assessment of multiple-choice items is conducted by an optical mark reader. The scores are added up to find a total score for each student for each part of the exam. Then, the total number of correct answers and the conversion of the scores to 100 are shared with students.

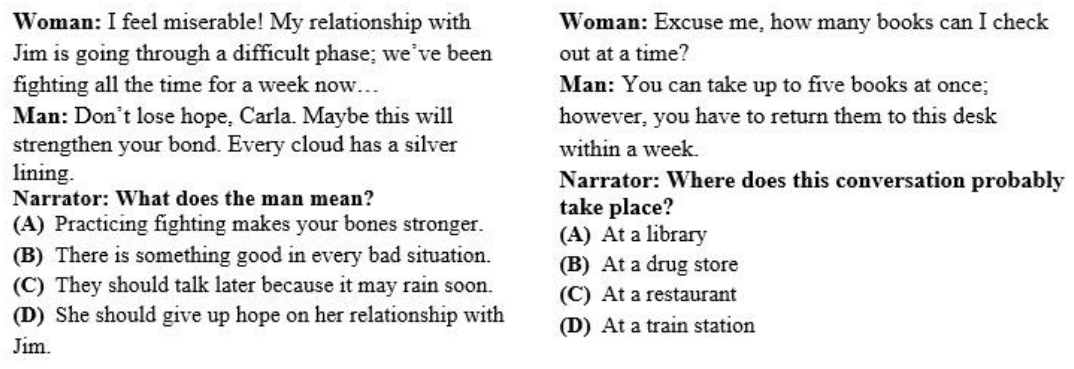
Marking sessions for writing exams are carried out by the instructors in the same room and at the same time. Instructors gather in the room and start marking session by grading a randomly selected student paper for standardization. The instructors are then paired to mark the exam packs and each student's paper is graded by two instructors. If the difference between the markers is more than 15 out of 100, they have to re-read the paper or ask measurement and evaluation unit members for their opinion.

As it can be seen in the table above (Table 1), in each midterm and quiz (except for the speaking quiz), there is also a section for vocabulary. However, vocabulary items were not examined in this study as it is not one of the main domains in English. In course objectives, there are no objectives specifically written for vocabulary. Furthermore, although writing is one of the main domains in English language, all writing parts in the exams have the same type of item which requires students to write essays. What's more, writing section cannot be coded with 0 and 1 and as a result of this, item parameters cannot be calculated. Thus, writing items were not examined in this study as well.

In this study, all quizzes and midterms (except vocabulary and writing parts) conducted in B level during the spring semester of 2016-2017 academic year were used. However, because of the security issues, all the exams and items could not be shared. Therefore, Figures 1, 2, 3 and 4 below show sample items from each domain.

Figure 1

Sample Listening Items

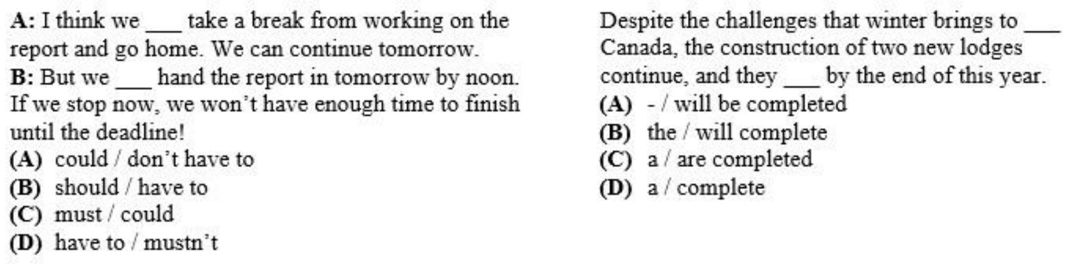


Woman: I feel miserable! My relationship with Jim is going through a difficult phase; we've been fighting all the time for a week now…
Man: Don't lose hope, Carla. Maybe this will strengthen your bond. Every cloud has a silver lining.
Narrator: What does the man mean?
(A) Practicing fighting makes your bones stronger.
(B) There is something good in every bad situation.
(C) They should talk later because it may rain soon.
(D) She should give up hope on her relationship with Jim.

Woman: Excuse me, how many books can I check out at a time?
Man: You can take up to five books at once; however, you have to return them to this desk within a week.
Narrator: Where does this conversation probably take place?
(A) At a library
(B) At a drug store
(C) At a restaurant
(D) At a train station

Figure 2

Sample Grammar Items

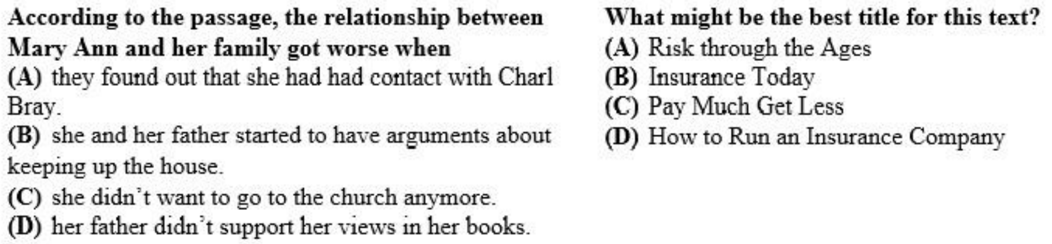


A: I think we ___ take a break from working on the report and go home. We can continue tomorrow.
B: But we ___ hand the report in tomorrow by noon. If we stop now, we won't have enough time to finish until the deadline!
(A) could / don't have to
(B) should / have to
(C) must / could
(D) have to / mustn't

Despite the challenges that winter brings to ___ Canada, the construction of two new lodges continue, and they ___ by the end of this year.
(A) - / will be completed
(B) the / will complete
(C) a / are completed
(D) a / complete

Figure 3

Sample Reading Items

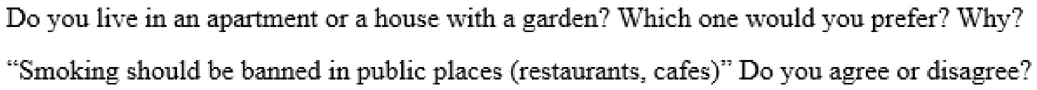


According to the passage, the relationship between Mary Ann and her family got worse when
(A) they found out that she had had contact with Charl Bray.
(B) she and her father started to have arguments about keeping up the house.
(C) she didn't want to go to the church anymore.
(D) her father didn't support her views in her books.

What might be the best title for this text?
(A) Risk through the Ages
(B) Insurance Today
(C) Pay Much Get Less
(D) How to Run an Insurance Company

Figure 4

Sample Speaking Items



Do you live in an apartment or a house with a garden? Which one would you prefer? Why?

"Smoking should be banned in public places (restaurants, cafes)" Do you agree or disagree?

Method of Data Collection

In data collection process, the syllabi of the courses, course book objectives and standardized classroom environment exams given to students were used. The syllabi of each course which were prepared by Curriculum Development Unit were first gathered. The exam data were collected from 458 students from B level. Although some students missed some of the exams due to various reasons such as having a sick report, missing data did not exceed 5% for each exam. The data of this study were collected throughout the spring semester of 2016-2017 academic year and this process lasted approximately three months.

For multiple-choice exam items, the students were given optic forms during the exams and they marked their answers on these forms. Then, these forms were transformed into excel files and the answers of test items were collected. While assessing speaking proficiency of students, the speaking exam of each student was recorded into the computer. For writing assessment, students were asked to write a paragraph or an essay and these papers were collected to be marked by the instructors based on the rubrics. In exam results obtained from Measurement and Evaluation Unit, there is no ID information as students' names were deleted. As I stated above, quizzes and midterms (except vocabulary and writing parts) conducted in B level were used.

Method of Data Analysis

The first step of data analysis was the content analysis of objectives and test items. At the beginning of this study, B level course objectives and exam items were taken from the institution.

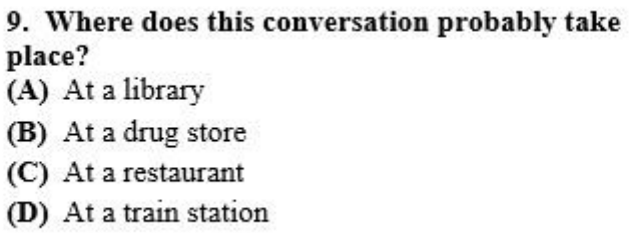
In order to determine the alignment level between the test items and learning objectives, instead of rating method, matching technique was used in this study and in this process some elements of content validity described by Sireci (1998) has been taken into consideration. In the first phase of this study, each test item was matched with learning objectives. This was done without any rating and if the item matched with objectives, the objective number was given and if not a dash (-) was used to indicate that there was no match.

This process was done by two raters for inter-rater agreement issues and before starting the matching process, forty items were chosen randomly from different exams and different domains to check the inter-rater agreement. These items were matched with objectives by two raters separately (one of the raters was a subject-matter expert). During this process, each item was taken separately and was matched with a learning objective. After matching, two raters checked their matchings and 85% agreement was found (Out of forty items, thirty-four items were matched with the same objectives and six items were matched with different objectives). Afterwards, raters discussed the items that they matched differently and any discrepancies were resolved by discussion, and finally 100% agreement was found. Here the inter-rater reliability was important for the correct matching of objectives and test items. After matching the sample forty items and checking the inter-rater agreement, all items were matched with learning objectives. As a result of this process, the agreement level of objectives and items was identified. The objectives which were not measured or the test items which were irrelevant (not matching with any objectives) were also identified.

To give an example for the matching process, please see Figure 5 and Figure 6 below.

Figure 5

Sample Listening Item Matching

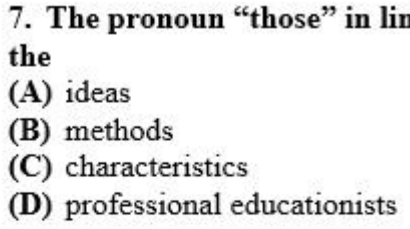


9. Where does this conversation probably take place?
(A) At a library
(B) At a drug store
(C) At a restaurant
(D) At a train station

When we look at the sample listening item that asks where the conversation takes place in Figure 5, this item was matched with two objectives. Students need to catch the details in the short conversation (CO1) and based on these details, they need to infer (CO4) where the conversation takes place. So, this item was matched with CO1 and CO4 in the listening section.

Figure 6

Sample Reading Item Matching



7. The pronoun "those" in line 5 refers to the

(A) ideas
(B) methods
(C) characteristics
(D) professional educationists

When we look at the sample reading item in Figure 7, the item asks for the pronoun reference. However, when we look at the course objectives, none of the objectives is about pronoun reference, so this item could not be matched. On the other hand, in course book objectives, CBO3 is about understanding pronoun reference; thus, although this item was not matched with any of the COs, it was matched with CBO3.

After participants' answers were uploaded into the computer through optic form reader, the results of each student were calculated via Excel.

Item analysis was conducted on SPSS. Based on the exam results, item difficulty and item discrimination values of each item were calculated. Item difficulty is the proportion of students who answer an item correctly to the total number of students. Difficulty values of items range between 0 and 1. Items which are closer to 0 are more difficult than those which are closer to 1. As for the item discrimination, the data were analyzed to determine the discriminating power of items and whether each item could discriminate the students or not. Each item is expected to discriminate between students who get high and low on a test. Item discrimination values range between -1 and 1. The higher the value is, the better the item is in terms of discrimination. In other words, item discrimination values which are close to 1 show that the item has a high discriminating power.

After the item analysis, questions were categorized as easy, mediocre or difficult depending on the item difficulty values and as discard, needs revision, mediocre, good and very good depending on the item discrimination values. These values and labels were taken from the institution and the same values and labels were used. Each value and level of difficulty and discrimination was supported with the literature. During the calculation of item parameters, Cronbach alpha values of each exam were also calculated on SPSS to provide an idea on reliability. Cronbach alpha values of quiz 1, quiz 2 and quiz 3 are .65, .58 and .71 respectively and .86, .86 and .87 for midterms. As speaking exams do not have multiple choice items and they cannot be coded with "0" and "1", item analysis could not be conducted for speaking quiz.

Table 2 below shows item difficulty and discrimination values and levels. They were set based on Ebel and Frisbie (1991) and taken from the institution.

Table 2

Item Difficulty and Discrimination Indices

Item Difficulty Value	Item Difficulty Label	Item Discrimination Value	Item Discrimination Label
0.00 - 0.29	Difficult	Below 0	Discard
0.30 - 0.79	Mediocre	0.00 - 0.19	Needs revision
0.80 - 1.00	Easy	0.20 - 0.29	Mediocre
		0.30 - 0.39	Good
		0.40 - 1.00	Very good

CHAPTER 4: RESULTS

Introduction

The purpose of this study was to determine whether there is an alignment between the learning objectives of courses taught and items in formative and summative assessment tools of English preparatory class in a higher education institution in Turkey. Another purpose of this study was to analyze the quality of assessment tools in terms of item difficulty and item discrimination.

In the first part of this chapter, results regarding the alignment between objectives and items were provided for each domain separately. For each domain, first quizzes and after quizzes midterms were examined. In the second part of the chapter, results of item analysis were given following the same pattern.

Match Between Learning Objectives and Items

Listening Domain

Table 3 below shows the listening course objectives (hereafter CO) and Table 4 shows course book objectives (hereafter CBO) and short phrases used to summarize them. As it is clearly seen from the Tables below (Table 3 & Table 4) listening COs mostly match with CBOs. Although the second objective focuses especially on long conversations and talks, listening CO1 and CO2 match with CBO2 as all of them focus on understanding specific details and information while the students are listening to a talk, conversation or lecture. Furthermore, objective 3 in both COs and CBOs are about understanding speaker’s tone and attitude. Similarly, the fourth objectives in both match as both objectives are about making inferences. Only the first objective of CBOs, which is understanding the main ideas, does not directly match any of the COs as there is no specific objective focusing on understanding the main ideas.

Table 3

Course Objectives and Phrases for Listening Domain

Course Objectives	Phrases
The students who successfully complete the listening & speaking class in the B level will be able to:	
1. understand factual information & specific details and identify general messages provided speech is clear and a generally familiar accent is used.	factual information & specific details
2. understand extended speech, long conversations and lectures and follow even complex lines of argument provided the topic is reasonably familiar	extended speech & lectures
3. understand most radio documentaries and most other recorded or broadcast audio, material delivered in standard language and identify the speaker’s mood, tone, etc.	recorded audio & speaker's tone and attitude
4. make deductions by using context clues	deduction

Table 4

Course Book Objectives and Phrases for Listening Domain

Course Book Objectives	Phrases
The students who successfully complete Pathways 2 in B level will be able to:	
1. understand main ideas of talks, lectures and dialogues	main ideas
2. understand details & specific information of talks and conversations about familiar topics	details & specific information
3. interpret speaker's tone and attitude	speaker's tone & attitude
4. make inferences out of context	inference

Quizzes

Table 5 shows COs and CBOs which were covered by items in the listening quizzes (quiz 1 and 3). Quiz 2 did not include listening domain, so it is not shown in Table 5.

Table 5

Listening Quiz Items

	Quiz 1		Quiz 3	
Items	CO	CBO	CO	CBO
Listening 1	4	4	1	2
Listening 2	4	4	1	2
Listening 3	1, 4	2, 4	1	2
Listening 4	4	4	1, 4	2, 4
Listening 5	1	2	1, 4	2, 4
Listening 6	1, 2	2	1, 4	2, 4
Listening 7	1, 2	2	2	1
Listening 8	1, 2	2	1, 2	2
Listening 9	1, 2	2	1, 2	2
Listening 10	1, 2	2	2, 4	4

Note. Each item in the first column corresponds to a specific item in each exam. The number of listening course objectives is 4.

Quiz 1 had 10 listening items in total. When we look at Table 5, most of the items (6 items) measured more than 1 CO. Six items measured 2 COs while 4 of them measured only 1. If we look at the distribution of COs, 7 out of 10 items measured CO1 (factual information and specific details), so we can say that CO1 is the most measured objective in quiz 1. Also, 40% of items required students to make some inferences (CO4) based on the context. Half of the items measured students' ability to understand long conversations or lectures (CO2). In these items, CO1 and CO2 were measured together as students needed to understand long talks or conversations and answered the detail questions. The Table also shows that CO3 was never measured in quiz 1 as there was no item regarding the tone or attitude of the speaker and students did not listen to recorded audio or broadcast either.

When we look at quiz 1 in terms of CBOs, unlike COs, almost all items except item 3 measured only 1 objective. That is because CO2 was specifically written for understanding long conversations and talks, but in CBOs, there is no such discrimination. According to Table 5, CBO number 2 (details and specific information) was the predominately measured objective as most items (8 out of 10) required students to understand details and specific information. Only item 3 measured 2 different objectives. Similar to COs, 40% of items measured the 4th CBO which is making inferences. It can be clearly seen from the Table that CBO1 and CBO3 were not measured in quiz 1 which means there was no question regarding the main idea or speaker's tone and attitude.

Quiz 3 also had 10 items. Table 5 indicates that 6 out of 10 items measured 2 COs while 4 of them measured only 1. Among the COs, CO1 (factual information & specific details), which was measured in 8 items, was the most frequently tested objective in quiz 3. Besides CO1, CO2 (extended speech and lectures) and CO4 (deduction) were also measured in 4 items. In this exam, 60% of items measured students' ability to understand short conversations whereas 40% of them measured the ability to understand a lecture. Similar to quiz 1, there was no item regarding the tone or attitude of the speaker as CO3 was not tested.

In terms of CBOs for quiz 3, CBO2 was tested almost in all items as the items mostly based on understanding specific details. Like CO4, which is about making deductions, CBO4 was also tested in 4 items. In quiz 3, %70 of items measured 1 CBO and the rest of the items measured more than 1 CBO. These 3 items measured students' ability to understand details and making inferences based on these details. Only item 7 measured CBO1 which is about understanding the main idea. As the last 4 items measured students' ability to understand a lecture, there was 1 item asking for the main idea of the lecture. In quiz 3, there was no item measuring the students' ability to understand speaker's tone and attitude as well.

If we take the quizzes as a whole, 12 items measured more than 1 CO and 40% of items measured only 1 CO. CO1 (factual information & specific details) was the most predominant objective tested in quizzes. Approximately half of the items measured CO2 (extended speech & lectures) and CO4 (deduction). However, none of the items in quizzes measured CO3 (recorded audio & speaker's tone and attitude).

As for CBOs, unlike COs, only 20% of items measured more than 1 objective and most of the items measured 1 CBO. Among the 20 items, CBO2 (details & specific information) was the most frequently tested CBO with the frequency of 15. Students' ability to make inferences (CBO4) was also measured in 40% of items. However, as the main focus was on understanding the details and making inferences, understanding the main idea skill was measured only in 1 item in both quizzes. As mentioned above, there was no item measuring interpretation of speaker's tone and attitude.

Midterms

Students' listening skills were also measured in midterm exams with more items. Table 6 below shows COs and CBOs covered in midterm exams.

Table 6

Listening Midterm Items

	Midterm 1		Midterm 2		Midterm 3	
Items	CO	CBO	CO	CBO	CO	CBO
Listening 1	1	2	1, 4	2, 4	1	2
Listening 2	1, 4	2, 4	1, 4	2, 4	1, 4	2, 4
Listening 3	1	2	1, 4	2, 4	1	2
Listening 4	1, 4	2, 4	1	2	1	2
Listening 5	1, 4	2, 4	1	2	1	2
Listening 6	1, 4	2, 4	1	2	1, 4	2, 4
Listening 7	1	2	1	2	1, 4	2, 4
Listening 8	1	2	1, 4	2, 4	1	2
Listening 9	1, 4	2, 4	1, 4	2, 4	1	2
Listening 10	1, 4	2, 4	1, 4	2, 4	1, 4	2, 4
Listening 11	1, 2	2	1	2	1	2
Listening 12	1, 2	2	1	2	1, 4	2, 4
Listening 13	1, 2	2	1, 2	2	2	1

Table 6 (cont'd)

Listening Midterm Items

Listening 14	1, 2	2	1, 2	2	1, 2	1, 2
Listening 15	1, 2	2	1, 2	2	1, 2	1, 2
Listening 16	1, 2	2	1, 2	2	2, 4	4
Listening 17	1, 2	2	2	1	2	1
Listening 18	1, 2	2	1, 2	2	1, 2	2
Listening 19	1, 2	2	1, 2	2	1, 2	2
Listening 20	1, 2	2	1, 2	2	2, 4	4
Listening 21	NA	NA	NA	NA	1, 3	2
Listening 22	NA	NA	NA	NA	1, 3	2
Listening 23	NA	NA	NA	NA	1, 3	2
Listening 24	NA	NA	NA	NA	1, 3	2
Listening 25	NA	NA	NA	NA	1, 3	2

Note. Each item in the first column corresponds to a specific item in each exam. Number of items in midterm 3 is different. Midterm 1 and 2 have 20 items while midterm 3 has 25. NA stands for not available. The number of listening course objectives is 4.

In midterm 1, 15 out of 20 items measured 2 COs and 5 of the items measured only 1 CO. All of the items in midterm 1 measured CO1 (understanding factual information and specific details). Besides CO1, 6 items measured CO4 (making deductions) and as 10 items were about a long conversation and a lecture, they also measured CO2. The only missing objective is CO3 (speaker's tone and attitude).

In terms of CBOs for midterm 1, 6 items measured more than 1 objective while 14 of them measured only 1. As CO1 and CBO2 measured the same skill, which is understanding specific details, all items measured CBO2 in midterm 1. As CBO4 and CO4 (deductions) also correspond, 6 of the items also measured CBO4. As there was no main idea item or tone or attitude question, CBO1 and CBO3 were not tested in midterm 1.

In midterm 2, we can observe a very similar Table with midterm 1 as CO1 (factual information & specific details) was measured almost in all items again. Only the last 8 items were about understanding the long conversation and lecture, so they also measured students' ability to understand them. In 30% of items students were expected to make inferences about what they listened to. However, understanding speaker's tone and attitude was not tested in midterm 2 as well.

As CO1 & CBO2 (factual information & specific details) and CO4 and CBO4 (deductions) correspond, these CBOs were measured in the same items with COs. When we look at the distribution of CBOs, we notice that there was just one item measuring students' ability to understand main ideas in the whole exam and all other items were about understanding the details and specific information.

In midterm 3, 16 out of 25 items measured more than 1 CO while the rest of them measured only 1. When we look at the distribution, approximately 80% of items in midterm 3 measured CO1. So, it is clear that most of the items in midterm 3 also measured students' ability to understand details and specific information. CO4 was the second mostly measured objective with the frequency of 7 out of 25. It suggests that a quarter of items measured the ability of making inference. In midterm 3, students also listened to a broadcast audio on a radio and last 5 items measured this ability. So, CO3 was partly tested for the first time in midterm 3.

When we look at CBOs for midterm 3, like other midterms, CBO2 was the most predominantly measured CBO as most items measured understanding details in short or long conversations, talks or recorded audios. Only 2 items measured students' ability to understand main ideas. As CO4 and CBO4 directly match (making inference), similar to CO4, CBO4 was also measured in 7 items.

If we have a look at midterms as a whole, a great majority of items (n=60) measured understanding details and specific information (CO1) both in short and long conversations and talks. Slightly above one third of the items (n=26) measured students' ability to understand long conversations or talks (CO2). The ability to make inference was also measured in 19 items out of 65. As for CBOs, CBO2 (details and specific information) was measured almost in all items. Similar to CO4, CBO4 (making inference) was also measured in 19 items. Based on the information provided above, we can say that midterm items mostly focused on understanding the details and specific information and making inferences. There was no item measuring the skill of understanding speaker's tone and attitude.

Grammar Domain

Table 7 below shows course objectives and phrases and Table 8 shows course book objectives and phrases for Grammar domain. When we look at the objectives, we may notice that COs are much more general when compared to CBOs. While writing the COs, students' use of language was taken into consideration in broader perspective. On the other hand, CBOs focused more on the books' objectives unit by unit as expected. Thus, it is not quite easy to match COs and CBOs because of this issue. In short, COs focus on more general objectives while CBOs focus more on specific ones in accordance with the course book.

Table 7

Course Objectives and Phrases for Grammar Domain

Course Objectives	Phrases
The goal of the main course in the B level is to make students reach the C level at the end of the term. The students who successfully complete the B level can:	
1. use appropriate language for different aims like persuading, informing or criticizing.	Using proper language for different aims
2. understand and produce enough to manage simple, routine tasks without delay	Understanding & producing simple tasks

Table 7 (cont'd)

Course Objectives and Phrases for Grammar Domain

3. form the necessary structure for asking for and following detailed directions	Forming the structure for directions
4. use proper structures to describe habits, routines or past experiences	Using proper structures to describe habits & experiences
5. cope with less routine situations in shops, post offices, banks, e.g. returning an unsatisfactory purchase	Dealing with routine situations
6. build the necessary structures for a complaint about everyday situations	Forming the structure for complaints
7. produce clear, well-structured sentences and texts on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.	Producing well-structured sentences & texts
8. forming sentences about unreal conditions and talk about their dreams and wishes	Forming sentences for unreal conditions
9. make his/her opinions and reactions understood about the question of what to do next, giving brief reasons and explanations, etc.	Giving reasons & explanations

Table 8

Course Book Objectives and Phrases for Grammar Domain

Course Book Objectives	Phrases
The students who successfully complete Language Leader Intermediate Level in B level can:	
1. form object & subject questions	Forming questions
2. use simple present & present continuous tenses accurately	Using simple present & present continuous tense correctly
3. distinguish between past tenses & present perfect tense	Using past tenses & present perfect tense correctly
4. use conditionals correctly (first-second-third)	Using conditionals correctly
5. make comparisons	Making comparisons
6. use future forms correctly	Using future forms correctly
7. form defining & non-defining relative clauses	Forming relative clauses
8. distinguish between modal verbs	Using modal verbs correctly
9. use articles accurately	Using articles correctly
10. distinguish between active and passive sentences	Forming passive & active sentences
11. use expressions of quantity correctly	Using expressions of quantity correctly
12. form reported speech	Forming reported speech
13. distinguish between ing form and to infinitive	Using ing form and to infinitive

Quizzes

Table 9 below shows quiz items and COs & CBOs covered in the quizzes. As stated in Chapter 3, quiz 1, quiz 2 and quiz 3 had grammar sections, but quiz 4 did not have a grammar section as it was only a speaking exam. There were 10 grammar items in each quiz.

Table 9

Grammar Quiz Items

	Quiz 1		Quiz 2		Quiz 3	
Items	CO	CBO	CO	CBO	CO	CBO
Grammar 1	1	1, 3	7	8	1, 8	8, 13
Grammar 2	1	3	7	3	7	11
Grammar 3	1	2	7	5	1	8
Grammar 4	1	3	7	7	1, 9	1, 7
Grammar 5	7	3	7	5	4	8
Grammar 6	1	2, 3	4	3	7	-
Grammar 7	7	1	1, 9	8	1	13
Grammar 8	7	2	-	9	8	4
Grammar 9	7	2, 3	1, 9	3, 8	1, 9	-
Grammar 10	7	1	7	3, 7	7	7

Note. Each item in the first column corresponds to a specific item in each exam. The number of grammar course objectives is 9.

In quiz 1, all of the items measured only 1 course objective. Moreover, only CO1 (using proper language for different aims) and CO7 (producing well-structured sentences & texts) were measured in the first quiz. Half of the items measured CO1 and the other half measured CO7. These two objectives were the most general ones and other COs were more theme specific. For example, some of them only focused on giving directions, making a complaint or describing habits and routines. CO1 (using proper language for different aims) and CO7 (producing well-structured sentences and texts) can be thus matched with most items in the exam.

As I stated above, CBOs are much more specific and they follow the course book’s units and were written accordingly. Thus, depending on the content of the exam, CBOs covered in the exams also changed. When we look at quiz 1, first 3 CBOs were covered in quiz 1 and 3 of the items measured more than 1 objective while 7 of them measured only 1. CBO3 (using past tenses & present perfect tense correctly) was the most frequently measured one with 6 items. CBO2 (using simple present & present continuous tense correctly) was measured in 4 items while CBO1 (forming questions) was measured in 3 items. In quiz 1, we can say that students' ability to use proper language and produce well-structured sentences were measured in terms of COs and their ability to form questions and use present and past tenses were measured in terms of CBOs.

In quiz 2, 20% of items measured more than 1 objective and 70% of them measured one. In quiz 2, we had an item which did not match any of the COs. CO7 (producing well-structured sentences & texts) was the most predominantly measured one with the frequency of 6 and 1 item measured CO4 (using proper structures to describe habits & experiences) while 2 of them measured CO9 (giving reasons & explanations). CO1 (using proper language for different aims) was also measured in 2 items.

When we take CBOs into consideration, we can observe more variety. We can see that 5 different CBOs were measured in 10 items. Only 1 item measured 2 different CBOs and 9 of them measured just 1 CBO. The distribution of CBOs was also quite acceptable as none of the CBOs dominated the exam. CBO3 (using past tenses & present perfect tense correctly) had the highest frequency and it was measured only in 4 items. Other CBOs were measured in 1 or 2 items mostly. Considering this data, we can say that students' ability to use past & perfect tenses correctly, make comparisons, form relative clauses and use articles were measured in quiz 2.

In quiz 3, 3 items measured 2 COs and 7 of them measured 1 CO. As in quiz 1 and 2, CO1 and CO7 were measured in this quiz again. In quiz 2, 7 out of 10 items either measured CO1 (using proper language for different aims) or CO7 (producing well-structured sentences & texts) along with other COs. CO4, CO8 and CO9 were the other COs measured in quiz 2 and each of them was measured either in 1 or 2 items.

When we look at CBOs, 2 items measured 2 CBOs, 6 items measured 1 CBO and 2 items did not match with any of the CBOs. The distribution is also quite good in terms of CBO as the highest frequency is just 3. It means the exam did not focus on just 1 CBO and there was variety again. However, there were also 2 items which did not match any of the CBOs. This means, in the exam there were 2 grammar items which were not covered in the course book. To summarize, in quiz 3 students' ability to form unreal conditions and relative clauses, use modal verbs and expressions of quantity and form sentences with gerund-infinitive were measured.

In all quizzes, there was an item (quiz 2, item 8) which did not match with any of the COs and there were 2 items (quiz 3 items 6, 9) which did not match any of the CBOs. Moreover, only 5 out of 30 items measured 2 COs and the rest of the items measured just 1 CO. In the same way, 7 out of 30 items measured 2 CBOs. In quizzes, CO1 (using proper language for different aims) and CO7 (producing well-structured sentences) were measured predominantly. However, almost half of the COs were never tested in quizzes. In terms of CBOs, there was much more variety when compared to COs. Out of 13 CBOs, 10 of them were measured in quizzes. Most frequently measured CBO was CBO3 (using past tenses and present perfect tense correctly) which was measured in 10 items. The distribution is also quiet good as nearly all other CBOs were measured only in 3 or 4 items except CBO3. This suggests that students' ability to use past tenses and perfect tenses were measured the most.

Midterms

Table 10 shows midterm items and COs & CBOs covered in the midterms. As stated in chapter 3, grammar section item numbers change in midterms. In midterm 1 and midterm 2 there were 20 grammar items while in midterm 3 there were 30.

Table 10

Grammar Midterm Items

	Midterm 1		Midterm 2		Midterm 3	
Items	CO	CBO	CO	CBO	CO	CBO
Grammar 1	7	3	7	3	7	9
Grammar 2	7	2	1, 9	11	7	3
Grammar 3	7	3	7	7	7	11
Grammar 4	7	7	1	13	7, 9	8
Grammar 5	7	3	1	13	7	13
Grammar 6	7	6	7	-	7	7
Grammar 7	1, 9	4	8	4	7	3
Grammar 8	7	3	7	-	7	3
Grammar 9	1, 7	7	1	5	1, 9	10
Grammar 10	8	4	1, 9	6	7	7
Grammar 11	8	1, 4	1, 9	8	1, 9	3
Grammar 12	1, 4	3	7	9, 10	1, 9	3
Grammar 13	9	2	7	8	4	2, 3
Grammar 14	7, 9	6	7, 9	9	7	6
Grammar 15	1	3	7	12	1	6, 10
Grammar 16	7	3	1	10	8	4
Grammar 17	1, 9	4	7	13	7	3, 4
Grammar 18	8	4	7	11	4	3, 8
Grammar 19	7	7	7	3	9	13
Grammar 20	7	2	7	13	7	3, 8
Grammar 21	NA	NA	NA	NA	4	8
Grammar 22	NA	NA	NA	NA	9	6
Grammar 23	NA	NA	NA	NA	7	-
Grammar 24	NA	NA	NA	NA	9	8
Grammar 25	NA	NA	NA	NA	4, 7	-
Grammar 26	NA	NA	NA	NA	7	7
Grammar 27	NA	NA	NA	NA	7	3

Table 10 (cont'd)

Grammar Midterm Items

Grammar 28	NA	NA	NA	NA	8	4
Grammar 29	NA	NA	NA	NA	4	3
Grammar 30	NA	NA	NA	NA	7	6

Note. Each item in the first column corresponds to a specific item in each exam. Number of items in midterm 3 is different. Midterm 1 and 2 have 20 items while midterm 3 has 30. NA stands for not available. The number of grammar course objectives is 9.

Midterm 1 had 20 items and 25 % of them measured 2 COs while 75% measured just 1. COs 1, 4, 7, 8 and 9 were measured in midterm 1, but the rest of the COs were not measured. CO7 (producing well-structured sentences) was measured in 60% of items and it was followed by CO1 (using proper language for different aims) which was measured in 5 items. Other COs measured in midterm 1 were tested in less than 5 items. When we look at CBOs, there was only 1 item that measured more than 1 objective, and other items measured only 1 objective. As CBOs are based on the book content, we again had more variety in terms of distribution. Objectives of all units which were covered until midterm 1 were tested in midterm 1. Most predominantly measured objective was CBO3 (using past tenses and present perfect tense correctly) with the frequency of 7 and all other objectives were measured less than 5 items.

In midterm 2, 4 items measured 2 COs and 16 items measured only 1 CO. The distribution of COs is quite similar to midterm 1 as expected, which means the main focus was on CO1 and CO7 again. 12 items measured CO7 and 7 items measured CO1. Besides CO1 and CO7, CO8 (forming sentences for unreal conditions) and CO9 (giving reasons and explanations) were also tested, but other COs were not measured. When we look at CBOs for midterm 2, only 1 out of 20 items measured more than 1 objective. We also had 2 items in this exam (item number 6 and number 8) which did not match any of the CBOs.

In midterm 3, there were 30 grammar items, which means there were 10 more items when compared to midterm 1 and midterm 2. Among these 30 items, only 5 of the items measured more than one CO. CO7 (producing well-structured sentences and texts) was measured almost in half of the items and it was the predominant one as in other midterms. CO9 (giving reasons and explanations) was measured in 6 items while CO1 (using proper language for different aims) was measured in 4 of them. Other COs that were measured in midterm 3 were CO4 (using proper structures to describe habits & experiences) and CO8 (forming sentences for unreal conditions) both of which were measured in 5 or fewer items. Thus, in midterm 3, students were measured for their ability to use proper language for different aims, produce well-structured sentences, give reasons and explanations and use structures to describe past experiences. For CBOs, 5 out of 30 items measured more than 1 CBO, 2 items did not match with any of the CBOs and the rest of the items measured 1 CBO. The variety of CBOs is quite good when compared to COs. Unlike COs, 10 different CBOs were measured in midterm 3. The most frequently measured CBO was objective 3 (using past tenses and present perfect correctly) with the frequency of 11. Other CBOs were measured with the frequency changing between 1 and 5.

In midterms as a whole, there were 70 grammar items in total. Among these 70 items, 20% of them measured 2 COs while 80% of them measured 1 CO. CO7 (producing well-structured sentences) was the most predominant one in midterms with the frequency of 41. There were also COs (CO2, CO3, CO5, CO6) which were not measured in midterms. CBOs had better distribution than COs. Seven items measured 2 CBOs and the rest of them measured 1 CBO. In contrast to COs, all

CBOs were measured in midterms although very few of them were measured just once. CBO3 (using past and present perfect tenses correctly) was measured the most with the frequency of 20. All other CBOs were measured less than 10. Moreover, 4 of the items did not match with any of the CBOs in total. In the light of the information given above, in midterms, students were measured for all the grammar points they learned.

Speaking Domain

Table 11 below shows course objectives and phrases and Table 12 shows course books objectives and phrases for speaking domain. Students were evaluated only once during the semester, so only quiz 4 included speaking assessment.

Table 11

Course Objectives and Phrases for Speaking Domain

Course Objectives	Phrases
The students who successfully complete the listening & speaking class in the B level will be able to:	
1. make short explanations about their ideas and preferences	Stating ideas & preferences
2. make short presentations about familiar topics with some preparation	making a presentation
3. take part in conversations on a familiar topic in a predictable context	Taking part in a conversation
4. describe their personality in a complicated way as well as ask questions for learning about the personality of others	Describing personality
5. make polite corrections in a dialogue using appropriate vocabulary and phrases	Making corrections
6. talk about similarities and differences of things, plans or people	Taking about similarities & differences
7. state ideas on basic contradictory issues	Stating ideas in contradictory issues
8. perform in short role plays and spontaneously use the basic and necessary structures to express ideas in a predictable context	Performing role plays
9. summarize and give opinion about a short story, article, discussion or documentary	Summarizing

Table 12

Course Book Objectives and Phrases for Speaking Domain

Course Book Objectives	Phrases
The students who successfully complete Pathways 2 in B level will be able to:	
1. give advice & make suggestions	Giving advice
2. explain causes and effects	Explaining cause & effect
3. speculate about the future	Speculation about future
4. discuss problems	Discussing problems
5. express opinions	Expressing opinions

In speaking quiz, students were asked approximately 10 to 15 questions to test their speaking skills. Table 13 below shows the speaking exam items and their matching with COs and CBOs.

Table 13

Speaking Quiz Items

	Quiz 4	
Items	CO	CBO
Speaking 1	3, 4	-
Speaking 2	1, 3, 4	5
Speaking 3	1, 3, 4	5
Speaking 4	1, 6	5
Speaking 5	1, 3	5
Speaking 6	1, 3	5
Speaking 7	1, 7	4, 5
Speaking 8	1	3, 5
Speaking 9	1	3, 5
Speaking 10	1, 6	5
Speaking 11	1	1, 5
Speaking 12	1, 7	4, 5
Speaking 13	1, 7	4, 5
Speaking 14	1, 7	4, 5
Speaking 15	1	5

Note. Each item in the first column corresponds to a specific item in each exam. The number of speaking course objectives is 9.

When we look at quiz 4, out of 15 items 2 of them measured 3 COs at the same time. That is because items 2 and 3 (items measuring 3 objectives) were about personality and they were asking for students' ideas on this issue. Furthermore, items 2 and 3 had follow-up questions, so the flow was more like a conversation between the teacher conducting the assessment and the student. That's why these two items covered 3 objectives. Furthermore, 9 items measured 2 objectives and 4 items measured only 1 objective. Thus, approximately 70% of items measured more than one objective. Nearly all items in quiz 4 measured CO1, which was a very general objective. It is about stating ideas and preferences and as almost all items asked students to express their ideas and preferences, it was the most dominant CO. Other COs that were measured in the exam were CO3 (taking part in a conversation), CO4 (describing personality), CO6 (talking about similarities and differences) and CO7 (stating ideas in contradictory issues). This means that 4 of the COs (CO2, CO5, CO8, CO9) were not measured in the speaking exam. As I mentioned above, CO1 (stating ideas and preferences) was the most dominant one and all other COs which were measured in the exam had the frequency of 5 at the most. The reason why CO1 was measured almost in all items was that all the items in the exam asked for students' ideas due to the exam format. During the exam, teachers asked questions and students answered them based on their ideas.

When we look at CBOs, 7 items measured 2 objectives at the same time, 7 items measured 1 objective and 1 item did not match with any of the CBOs. CBO5 (expressing opinions) and CO1 (stating ideas and preferences) correspond with each other as both objectives are based on expressing ideas. As a result of this, similar to CO1, CBO5 was measured almost in all items. Except for CBO2, all other CBOs were measured in the exam and all of them had the frequency of 5 or less. To put it briefly, students were measured for most of COs and CBOs.

Reading Domain

Table 14 below shows course objectives and phrases for Reading domain and Table 15 shows course book objectives and phrases. When we compare COs and CBOs, we can clearly see that some objectives directly match with each other. To exemplify, CO3 and CBO1 are about skimming strategies and CO4 and CBO4 are about scanning strategies. Similarly, CO5 and CBO6 are about measuring the vocabulary knowledge of students. Moreover, CO7 and CBO5, which are about distinguishing facts from theories, match with each other. On the other hand, there are also some objectives which do not match such as understanding cause & effect relationship, summarizing, pronoun references and making inferences.

Table 14

Course Objectives and Phrases for Reading Domain

Course Objectives	Phrases
The students who successfully complete the reading & writing class in the B level will be able to:	
1. have a good understanding of the relationship between sentences and paragraphs	sentence-paragraph relationship
2. use a monolingual and/or bilingual dictionary to extend their vocabulary	using dictionary
3. skim text (that includes a relatively wide vocabulary range) in order to identify the purpose and main idea of the text	skimming
4. scan texts in order to locate desired information, and gather information from different parts of text, or from different texts in order to fill in a chart	scanning
5. work out the meaning of unknown words in a familiar context	guessing the meaning
6. understand the organization of a complicated text	understanding the organization of texts
7. differentiate between facts and opinions in texts with a wide vocabulary range	differentiating between facts and opinions

Table 15

Course Book Objectives and Phrases for Reading Domain

Course Book Objectives	Phrases
The students who successfully complete Reading Explorer 2 in B level can:	
1. skim for the main idea of paragraphs	Skimming
2. identify the purpose of paragraphs	Identifying the purpose
3. understand pronoun reference	Pronoun reference
4. scan for details	Scanning
5. distinguish facts from theories	Distinguishing Facts and Theories
6. deal with unfamiliar vocabulary	Guessing the meaning
7. differentiate between the main ideas and supporting details	Differentiating main ideas and details
8. understand cause & effect relationships	Understanding cause & effect
9. understand synonyms	Finding synonyms
10. understand inference	Understanding inferences
11. identify an author’s tone or purpose	Identifying author's tone & purpose
12. summarize a text	Summarizing

Quizzes

Table 16 below shows quiz items and COs & CBOs covered in quizzes.

Table 16

Reading Quiz Items

	Quiz 1		Quiz 3	
Item Numbers	CO	REO	CO	REO
Reading 1	1	10	3	1, 7
Reading 2	5	6	-	3
Reading 3	4	4	4	4
Reading 4	-	3	4	4
Reading 5	4	4	5	6
Reading 6	4	4	4	4
Reading 7	5	6	4	4
Reading 8	4	4	4	4
Reading 9	4	4	5	6
Reading 10	-	-	4	4

Note. Each item in the first column corresponds to a specific item in each exam.

The number of reading course objectives is 9.

In quiz 1, there was no item measuring more than 1 CO. There were 2 items which did not match with any of the COs and the rest of the items measured 1 CO. CO4, which was about scanning skills, was the objective which was measured the most in quiz 1. Half of the items in quiz 1 measured CO4 which means half of the items measured students' scanning skills. Other than CO4, CO5 was measured twice and CO1 was measured in 1 item. In terms of CBOs, similar to COs, there was no item measuring more than 1 CBO. 90% of items measured 1 CBO and there was 1 item which did not match with any of the CBOs. As CO4 and CBO4 directly match with each other as both of them focus on scanning skills, CBO4 was the most dominant objective as well. Like CO4, half of the items measured CBO4. Other CBOs measured in the exam were CBO3 (pronoun reference), CBO6 (guessing the meaning) and CBO10 (understanding inferences).

When we look at quiz 3, we had 1 item that did not match with any of the COs, and the rest of the items measured just 1 CO. Similar to quiz 1, in this exam CO4 (scanning) was the most dominant CO with the frequency of 6. Besides CO4, CO3 (skimming) and CO5 (guessing the meaning) were the other objectives measured in quiz 2. When we look at CBOs, the first item measured 2 objectives at the same time while other items measured just 1. CBO4 (scanning) was measured in 6 items similar to CO4 and other CBOs measured in quiz 2 were CBO1 (skimming), CBO3 (pronoun reference), CBO6 (guessing the meaning) and CBO7 (differentiating main ideas and details).

When we consider quiz 1 and quiz 3, 17 out of 20 COs measured 1 objective and none of the items measured more than 1 CO. Three items did not match with any of the COs. Approximately 50% of items measured CO4 which was based on students' scanning skills. Besides students' scanning skills, they were also tested for their skimming skills and vocabulary knowledge. In terms of CBOs, 1 item measured 2 objectives, 1 item did not match with any objectives and the rest of the items measured 1 objective. As CO4 and CBO4 match with each other (both of them measure scanning skills), similar to CO4, CBO4 was also measured approximately in half of the items. Other CBOs measured in quizzes were CBO1, CBO6, CBO7 and CBO10 which are about skimming skills, vocabulary knowledge and making inferences.

Midterms

Table 17 below shows COs and CBOs covered in midterm exams. In all midterm exams, there was a reading section and while midterm 1 and midterm 2 had 20 items each, midterm 3 had 25 items.

Table 17

Reading Midterm Items

	Midterm 1		Midterm 2		Midterm 3	
Items	CO	CBO	CO	CBO	CO	CBO
Reading 1	-	3	3	1, 11	1, 6	-
Reading 2	4	4	1, 4	4, 8	1, 6	-
Reading 3	5	6	-	3	1, 6	-
Reading 4	4	4	4	4, 10	1, 6	-
Reading 5	4	4	5	6	1, 6	-
Reading 6	3	1, 7	5	6	3	1, 11
Reading 7	-	3	4	4	4	4
Reading 8	4	4	4	4	-	3
Reading 9	5	6	4	4	4	4
Reading 10	4	4	4	4, 10	4	4
Reading 11	3	1, 7	3	1, 7	-	3
Reading 12	4	4	4	4, 10	4	4
Reading 13	1, 4	4, 8	-	3	4	4
Reading 14	4	4	5	6	4	4
Reading 15	-	3	4	4	6	7, 10
Reading 16	5	6	4	4	3	1, 7
Reading 17	4	4	4	4	-	3
Reading 18	-	3	4	4, 10	-	3
Reading 19	5	6	5	6	4	4
Reading 20	4	4	4	4	5	6

Table 17 (cont'd)

Reading Midterm Items

Reading 21	NA	NA	NA	NA	4	4
Reading 22	NA	NA	NA	NA	5	6
Reading 23	NA	NA	NA	NA	4	4, 10
Reading 24	NA	NA	NA	NA	4	4
Reading 25	NA	NA	NA	NA	-	10

Note. Each item in the first column corresponds to a specific item in each exam. Number of items in midterm 3 is different. Midterm 1 and 2 have 20 items while midterm 3 has 25. NA stands for not available. The number of reading course objectives is 9.

When we look at COs in midterm 1, only 1 item measured 2 objectives and 15 items measured 1 CO. 4 of the items did not match with any of the COs. Half of the items measured CO4 (scanning) and this objective was the most dominant one similar to quizzes. Other objectives measured in midterm 1 were CO1 (sentence-paragraph relationship), CO3 (skimming) and CO5 (guessing the meaning). For CBOs, 3 of the items measured 2 objectives while 17 of them measured just 1. Unlike COs, all items matched with CBOs in midterm 1. As CO4 and CBO4 both measure scanning skills, half of the items also measured CBO4. CBO1, CBO3, CBO6, CBO7 and CBO8 were also measured in midterm one although the highest frequency for these objectives was 4.

In midterm 2, only 1 item measured 2 objectives and 2 items did not match with any of the objectives. The rest of the items measured 1 CO. 12 out of 20 items measured CO4 (scanning) and it was the most dominant objective as in midterm 1. Other COs measured in midterm 2 were CO1 (sentence-paragraph relationship), CO3 (skimming) and CO5 (guessing the meaning). Other COs were not tested in midterm 2. When we look at CBOs, 7 items measured 2 objectives while 13 of them measured just 1 CBO. As CO4 and CBO4 were about scanning skills, similar to CO4, 12 of the items measured CBO4 as well and thus, it was the most dominant CBO. CBO1, CBO3, CBO6, CBO7, CBO8, CBO10 and CBO11 were also measured in midterm 2. However, CBO8 (understanding cause and effect) and CBO11 (summarizing) were measured only once in this exam.

In midterm 3, there were 25 items in the exam, which means there were 5 extra items when compared to midterm 1 and midterm 2. In this exam, there were 5 items measuring 2 objectives, 5 items which did not match with any COs and 15 items that measured only 1 CO. Similar to other midterms, CO4 (scanning) was once more the most dominant one with the frequency of 10. CO6 (understanding the organization of texts) comes the second with the frequency of 6 and CO1 (sentence-paragraph relationship) follows CO6 with the frequency of 5. Other than these COs, CO3 (skimming) and CO5 (guessing the meaning), which had the frequency of 2, were also measured. In terms of CBOs, 4 of the items measured 2 objectives, 5 items did not match with any CBOs and 16 items measured only 1 objective. CBO4 (scanning) was measured in 10 items similar to CO4. Other CBOs measured in midterm 3 were CBO1, CBO3, CBO6, CBO7, CBO10 and CBO11. Among these, CBO3 (pronoun reference) had the highest frequency of 4.

When we look at the reading items in midterms as a whole, it is clearly seen that CO4, which is about students' scanning skills, was the most dominant one in all midterm exams. Almost half of the items in midterms measured CO4. We can see that CO2 (using dictionary) and CO7 (differentiating between facts and opinions) were not measured in midterm exams. Furthermore, 11 items did not match with any of the COs. In terms of CBOs, similar to CO4, almost half of the items measured CBO4. Moreover, out of 65 items, only 5 items did not march with any of the CBOs.

Item Analysis

In this part of the chapter, results of the item parameter analyses were given. Analyses were presented based on domains. In this section, only listening, grammar and reading item analyses were provided. In each Table, item difficulty and item discrimination values and levels for each item were shown. Similar to item-objective matching part, first quiz item parameters and then midterm item parameters were provided. Items were labelled for both item difficulty and discrimination. See Chapter 3 for definition of labels.

Listening Domain

Quizzes

Table 18 below shows item difficulty and discrimination values and labels of listening quiz items.

Table 18

Listening Item Analysis (Quizzes)

	Difficulty		Discrimination		Difficulty Label		Discrimination Label	
Items	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
Listening 1	.99	.78	.01	.16	E	M	NR	NR
Listening 2	.95	.82	.06	.22	E	E	NR	M
Listening 3	.97	.50	.25	.23	E	M	M	M
Listening 4	.92	.86	.07	.28	E	E	NR	M
Listening 5	.87	.95	.26	.17	E	E	M	NR
Listening 6	.93	.92	.12	.26	E	E	NR	M
Listening 7	.98	.88	.13	.07	E	E	NR	NR
Listening 8	.95	.56	.16	.24	E	M	NR	M
Listening 9	.14	.89	.14	.16	D	E	NR	NR
Listening 10	.63	.91	.20	.20	M	E	M	M

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, D stands for difficult, M stands for mediocre, NR stands for needs revision.

In listening quiz 1, difficulties of items range between .14 and .99. Item 9 has the lowest difficulty (the hardest item) and item 1 (the easiest item) has the highest difficulty value. Median for item difficulty is .94. When we look at the values for quiz 1, only item 9 is below .30 and item 10 is between .30 and .80. 8 out of 10 items are above .80 which means that 80% of listening items in quiz 1 are classified as easy while 1 item is classified as difficult and 1 item as mediocre.

When we look at item discrimination values of quiz 1 listening items, they range between .01 for item 1 and .26 for item 5 (median .13). These values suggest that 7 items need revision as their values are below .20. There are 3 mediocre items in this exam whose values are between .20 and .30.

In quiz 3, listening item 3 (.50) and item 5 (.95) have the lowest and highest item difficulty values (median=.87). When we look at the values for quiz 3, there is no item whose value is below .30 and this suggests that there is no item classified as difficult in this exam. There are 3 items with the values between .50 and .78 and we can see that these items are considered as mediocre. 7 out of 10 items are above .80 which means that 70% of listening items in quiz 3 are classified as easy.

In terms of item discrimination values of quiz 3 listening items, they range between .07 (item 7) and .28 (item 4) with a median of .21. Based on these values, 4 items need revision as their values are below .20. The rest of the items’ values are between .20 and .30 and these items are classified as mediocre.

It is clearly seen that the difficulty levels of both quizzes are low as most of the items are labelled as easy. Similarly, discrimination levels of these exams show that almost half of the items need revision in order to discriminate better among students. To summarize, both the difficulty levels and discrimination levels of items need some revision.

Midterms

Item difficulty and discrimination values and labels of listening midterm items are shown in Table 19 below.

Table 19

Listening Item Analysis (Midterms)

	Difficulty			Discrimination			Difficulty Label			Discrimination Label		
Items	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Listening 1	.76	.94	.94	.38	.16	.16	M	E	E	G	NR	NR
Listening 2	.51	.79	.83	.33	.41	.09	M	M	E	G	VG	NR
Listening 3	.71	.86	.96	.39	.33	.14	M	E	E	G	G	NR
Listening 4	.80	.64	.42	.32	.38	.09	E	M	M	G	G	NR
Listening 5	.75	.87	.94	.38	.32	.14	M	E	E	G	G	NR
Listening 6	.93	.81	.95	.30	.17	.14	E	E	E	G	NR	NR
Listening 7	.75	.97	.70	.44	.17	.29	M	E	M	VG	NR	M
Listening 8	.73	.88	.95	.31	.11	.18	M	E	E	G	NR	NR
Listening 9	.92	.99	.62	.27	.03	.19	E	E	M	M	NR	NR
Listening 10	.96	.79	.89	.16	.27	.28	E	M	E	NR	M	M
Listening 11	.88	.93	.34	.24	.17	.31	E	E	M	M	NR	G
Listening 12	.92	.92	.98	.12	.24	.09	E	E	E	NR	M	NR
Listening 13	.36	.83	.87	.24	.29	.25	M	E	E	M	M	M
Listening 14	.46	.84	.89	.32	.24	.11	M	E	E	G	M	NR
Listening 15	.62	.69	.72	.28	.39	.28	M	M	M	M	G	M
Listening 16	.79	.77	.59	.17	.31	.31	M	M	M	NR	G	G
Listening 17	.75	.66	.64	.30	.20	.29	M	M	M	G	M	M
Listening 18	.90	.50	.56	.29	.35	.22	E	M	M	M	G	M
Listening 19	.52	.89	.16	.25	.01	.12	M	E	D	M	NR	NR
Listening 20	.89	.69	.62	.28	.31	.16	E	M	M	M	G	NR
Listening 21	NA	NA	.87	NA	NA	.19	NA	NA	E	NA	NA	NR
Listening 22	NA	NA	.99	NA	NA	.01	NA	NA	E	NA	NA	NR
Listening 23	NA	NA	.78	NA	NA	.16	NA	NA	M	NA	NA	NR
Listening 24	NA	NA	.99	NA	NA	.01	NA	NA	E	NA	NA	NR
Listening 25	NA	NA	.44	NA	NA	.18	NA	NA	M	NA	NA	NR

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, M stands for mediocre, D stands for difficult, NR stands for needs revision, G stands for good and VG stands for very good. NA stands for not available.

In midterm 1, difficulties of items range between .36 and .96. Item 13 has the lowest difficulty while item 10 has the highest difficulty value. Median for item difficulty is .75. When we look at the values for midterm 1, there is no item whose value is below .30 and this suggests that none of the items is classified as difficult in this exam. Twelve out of 20 items have values between .36 and .79 and these items are considered as mediocre. Eight items have values above .80 which means that 40% of listening items in midterm 1 are classified as easy.

Item 12 has the value of .12 and it is the weakest item for discrimination in midterm 1. On the other hand, item 7 is the best item with the value .44. The median for midterm 1 is .29. Thus, we can say that midterm 1 has 3 items that need revision, 7 mediocre items (between .24 and .29), 9 items with good discriminating power (between .30 and .39), and only 1 very good item (.44).

Most of the listening items are easy in midterm 2 and there is no difficult item. Even the most difficult item (item 18) has the value of .50 which means the most difficult item is answered correctly by half of the students. On the contrary, item 9 (.99) is the easiest item and almost all students answered this item correctly. Even the median value (.83) is very high and it is another sign that listening items are easy. Distribution of item difficult is 60% for easy items and 40% for mediocre items.

Although at least half of the students answered each item correctly in midterm 2, item discrimination values are still at acceptable levels (median=.25). The best item to discriminate among students is item 2 with the value of .41 and item 19 does not work well for discrimination (.01). Despite the fact that there is no difficult item, there are 8 items which are either good or very good for discrimination. Other than these 8 items, midterm 2 has 7 item that need revision and 5 mediocre items.

In midterm 3, unlike midterm 1 and 2, we have items from each difficulty level. The only difficult item of midterms is item 19 (.16) and it is the most difficult item of all midterms. Item 24 (.99) is the easiest item as almost all students answered it correctly. The median can also be considered high with the value of .83. Except for the difficult item, 11 out of 25 items (values between .34 and .78) are mediocre and 13 items are classified as easy.

Item discrimination values of midterm 3 show that items do not work well in general. The number of items that need revision can be considered high (n=17). As the number of items that need revision is a little high, the median for discrimination values is also not very high (.16). The values range between .01 (item 22 & item 24) and .31 (item 11 & item 16). Besides 17 items that need revision, there are 6 mediocre items (values between .22 and .29) and 2 items with good discriminating power.

When we take all the midterms into consideration, except for 1 difficult item, all the items are either mediocre or easy. So, item difficulty for midterms is a little low. Approximately half of the items are easy and the other half is mediocre. For item discrimination values, it can be seen that just below 50% of items need revision, so there are many items that do not work well for item discrimination. Only 2 items are classified as very good with values over .40. The rest of the items have equal number of mediocre and good items in terms of item discrimination.

If we are to compare item difficulty and discrimination values and labels, there are 38 listening items that need revision for item discrimination in both quizzes and midterms and among these items, 2 of them are difficult, 7 of them are mediocre and 29 of these items are easy. Out of 48 easy items, 29 of them need revision, 15 of them are mediocre and 4 are good items in terms of discrimination. As for mediocre items in terms of difficulty, there are 14 good, 12 mediocre and 2 very good items while only 7 of them need revision. So, mostly the easy listening items need revision.

Grammar Domain

Quizzes

For item difficulty and discrimination values and labels of grammar quiz items, see Table 20 below.

Table 20

Grammar Item Analysis (Quizzes)

	Difficulty			Discrimination			Difficulty Label			Discrimination Label		
Items	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Grammar 1	.94	.89	.66	.10	.14	.06	E	E	M	NR	NR	NR
Grammar 2	.98	.63	.44	.10	.29	.24	E	M	M	NR	M	M
Grammar 3	.64	.98	.76	.37	.03	.15	M	E	M	G	NR	M
Grammar 4	.92	.60	.44	.11	.22	.19	E	M	M	NR	M	NR
Grammar 5	.94	.52	.40	.14	.26	.09	E	M	M	NR	M	NR
Grammar 6	.83	.68	.64	.20	.35	.23	E	M	M	M	G	NR
Grammar 7	.62	.88	.76	.24	.20	.14	M	E	M	M	M	M
Grammar 8	.83	.91	.70	.23	.17	.17	E	E	M	M	NR	NR
Grammar 9	.99	.68	.79	.04	.11	.24	E	M	M	NR	NR	NR
Grammar 10	.96	.58	.66	.13	.36	.23	E	M	M	NR	G	M

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, M stands for mediocre, NR stands for needs revision and G stands for good.

In quiz 1, difficulties of grammar items range between .62 and .99. Item 7 has the lowest difficulty and item 9 has the highest difficulty. Median for item difficulty is .93. In quiz 1, none of the items is below .30. So, there is no item classified as difficult. Two out of 10 items are above .30 and below .80 which means that %20 of grammar items in quiz 1 are classified as mediocre. The difficulty levels for the rest of the items are easy their values are above .80.

When we look at discrimination values of quiz 1 grammar items, item 3 has the highest value (.37) while item 9 has the lowest (.04) with a median of .13. Based on these values, 6 items need revision as their values are below .20. There are 3 items whose values are between .20 and .29, so they are classified as mediocre. Only 1 item can be categorized as good as its value is higher than .30.

Even the most difficult grammar item of quiz 2 was answered by 52% of students which means .52 (item 5) is the lowest value. This also means that there is no difficult item in this exam similar to quiz 1. Moreover, .98 for item 3 is the highest value and this is the item almost all students answered correctly. (median .68). With no difficult item, quiz 2 has 6 mediocre (values between .52 and .68) and 4 easy items.

In quiz 2, whose median for item discrimination is .21, there are 4 items which need revision and the lowest value of these items is .03 (item 3). Four of the items are mediocre and there are also 2 items in quiz 2 (item 6 & item 10) and they are good items for discrimination with values above .30.

Quiz 3 (median=.66) is different from both quiz 1 and 2. Unlike other quizzes, all the items in quiz 3 have the same difficulty label. All difficulty values range between .40 (item 5, the most difficult) and .79 (item 9, the easiest), so all the items are mediocre in terms of difficulty.

All items in quiz 3 are either mediocre or need revision in terms of item discrimination. There are 6 items which need revision with values lower than .20 and 4 mediocre items (median=.18). The best item in quiz 3 is item 9 (.24) and the weakest item for discrimination is item 1 (.06).

When we look at the quizzes as a whole, none of the grammar quiz items is classified as difficult as there is no value below .30. Slightly more than half of the items (18) are considered as mediocre and the rest are easy items. Data show that difficulty levels of quizzes are not high. Discrimination values also show that items need revision to have more discriminating power as almost half of the items are classified as needs revision and one third as mediocre. Only 10% of items are classified as good which needs to be higher.

Midterms

Table 21 below shows items difficulty and discrimination values and labels for grammar midterm items.

Table 21

Grammar Item Analysis (Midterms)

	Difficulty			Discrimination			Difficulty Label			Discrimination Label		
Items	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Grammar 1	.78	.87	.71	.27	.27	.08	M	E	M	M	M	NR
Grammar 2	.66	.98	.70	.24	.04	.26	M	E	M	M	NR	M
Grammar 3	.97	.90	.77	.17	.27	.16	E	E	M	NR	M	NR
Grammar 4	.95	.68	.82	.25	.17	.29	E	M	E	M	NR	M
Grammar 5	.98	.98	.88	.18	.12	.19	E	E	E	NR	NR	NR
Grammar 6	.97	.63	.50	.14	.18	.38	E	M	M	NR	NR	G
Grammar 7	.87	.89	.59	.20	.20	.27	E	E	M	M	M	M
Grammar 8	.82	.93	.68	.25	.19	.26	E	E	M	M	NR	M
Grammar 9	.86	.79	.61	.36	.32	.34	E	M	M	G	G	G
Grammar 10	.82	.96	.41	.19	.20	.14	E	E	M	NR	M	NR
Grammar 11	.76	.90	.75	.34	.22	.34	M	E	M	G	M	G
Grammar 12	.90	.86	.74	.30	.32	.27	E	E	M	G	G	M
Grammar 13	.75	.47	.71	.27	.28	.32	M	M	M	M	M	G
Grammar 14	.72	.59	.91	.17	.22	.26	M	M	E	NR	M	M
Grammar 15	.85	.58	.51	.31	.24	.43	E	M	M	G	M	VG
Grammar 16	.92	.68	.68	.17	.12	.31	E	M	M	NR	NR	G
Grammar 17	.90	.90	.76	.16	.14	.31	E	E	M	NR	NR	G
Grammar 18	.70	.72	.84	.31	.28	.26	M	M	E	G	M	M
Grammar 19	.83	.60	.70	.29	.17	.20	E	M	M	M	NR	M
Grammar 20	.94	.52	.90	.26	.23	.29	E	M	E	M	M	M
Grammar 21	NA	NA	.38	NA	NA	.17	NA	NA	M	NA	NA	NR
Grammar 22	NA	NA	.69	NA	NA	.39	NA	NA	M	NA	NA	G
Grammar 23	NA	NA	.61	NA	NA	.29	NA	NA	M	NA	NA	M
Grammar 24	NA	NA	.82	NA	NA	.24	NA	NA	E	NA	NA	M
Grammar 25	NA	NA	.36	NA	NA	.18	NA	NA	M	NA	NA	NR

Table 21 (cont'd)

Grammar Item Analysis (Midterms)

Grammar 26	NA	NA	.80	NA	NA	.25	NA	NA	E	NA	NA	M
Grammar 27	NA	NA	.60	NA	NA	.30	NA	NA	M	NA	NA	G
Grammar 28	NA	NA	.53	NA	NA	.33	NA	NA	M	NA	NA	G
Grammar 29	NA	NA	.58	NA	NA	.28	NA	NA	M	NA	NA	M
Grammar 30	NA	NA	.75	NA	NA	.31	NA	NA	M	NA	NA	G

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, M stands for mediocre, NR stands for needs revision, G stands for good and VG stands for very good. NA stands for not available.

Item difficulty values of midterm 1 range between .66 and .98. Item 2 has the lowest value and 5 has the highest. Median for difficulty values of midterm 1 is .86. When we look at the values, there is no item with the value lower than .30, which means none of the items in midterm 1 is classified as easy. Six out of 20 items in midterm 1 have difficulty values between .66 and .78, so there are 6 mediocre items in this exam. The rest of the items which are equal to 70% are classified as easy since their difficulty values are above .80.

Item 9 is the most powerful item for item discrimination with the value .36. On the contrary, the weakest item is item 6 (.14). The median of item discrimination for midterm 1 is .25. Midterm 1 has 7 items that need revision (values lower than .19), 8 mediocre items (between .20 and .29) and 5 good items (values over .30).

Half of the items in midterm 2 are easy items while the other half are mediocre (median=.83). The most difficult item (.47) is item 13 and this item was approximately answered correctly by half of the students. On the other hand, almost all students answered item 5 correctly which makes it the easiest item (.98). Similar to midterm 1, there is no difficult item.

Half of the items in midterm 2 (median .21) are mediocre (between .20 and .28) in terms of item discrimination and the other half mostly need revision (40%). The best item to discriminate between high achievers and low achievers is item 12 (.32). The weakest item which does not work well for discrimination is item number 2 with the value of .04.

Midterm 3 (median=.70) has 30 items unlike midterm 1 and 2 and like midterm 1 and 2, we do not have any difficult item in this exam. Approximately one third of students answered correctly the most difficult item of the exam (item 25, .36). On the contrary, the easiest item of midterm 3 was answered correctly by 91% of students (item 14). A great majority of items (n=23) are mediocre items (values between .36 and .77). The rest of the items (n=7) are easy with values higher than .80.

Item discrimination values of midterm 3 are quite acceptable with the median of .28. Item 15 is the best item discriminating among students (.43) and the weakest one is item 1 with .08. Unlike most other exams, discriminating power of items can be considered much better in midterm 3. There are 13 mediocre items and 33% of items are good while only 20% of items need revision. There is also 1 very good item.

When we look at all the grammar items in midterms, there is no difficult item as the lowest value for item difficulty is .36. All the items in midterms are either mediocre (n=39) or easy (n=41). So, although the number of mediocre items is not very low or high, the difficulty of midterms can be a little higher. For item discrimination, most items are mediocre and good items, so discrimination seems acceptable although there are items which need revision.

Taking the item difficulty values and discrimination labels into consideration, we can see that out of 57 mediocre items in terms of difficulty, 16 of them are good, 23 mediocre, 17 need revision and 1 very good in terms of item discrimination. For easy grammar items, there are 4 good and 19 mediocre items and 20 of them need revision. So, it is clearly seen that while almost one third of mediocre items need revision, this ratio is one in two for easy items. As a result, easy items need to be revised for better discrimination.

Reading Domain

Quizzes

Item difficulty and discrimination values and labels for reading quizzes are given in Table 22.

Table 22

Reading Item Analysis (Quizzes)

	Difficulty		Discrimination		Difficulty Label		Discrimination Label	
Items	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
Reading 1	.92	.80	.14	.16	E	E	NR	NR
Reading 2	.73	.77	.07	.24	M	M	NR	M
Reading 3	.42	.85	.30	.18	M	E	G	NR
Reading 4	.90	.15	.10	.24	E	D	NR	M
Reading 5	.75	.68	.32	.18	M	M	G	NR
Reading 6	.83	.39	.28	.24	E	M	M	M
Reading 7	.48	.63	.30	.28	M	M	G	M
Reading 8	.93	.76	.22	.14	E	M	M	NR
Reading 9	.60	.62	.01	-.03	M	M	NR	DIS
Reading 10	.99	.68	.03	.29	E	M	NR	M

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, M stands for mediocre, NR stands for needs revision, G stands for good and DIS stands for discard.

Item difficulty values for reading items in quiz 1 range between .42 and .99. Item 3 has the lowest and 10 has the highest values. Median for item difficulty for quiz 1 is .79. In quiz 1, there is no difficult items as all the difficulty values are above

.30. Half of the items have values between .42 and .75. The other half of the items have values above .80. These values suggest that 50% of items are classified as mediocre and half of them as easy items.

Half of the reading items in quiz 1 need revision and among these item 9 is the weakest item with .01. There are 2 items with good discriminating power and item 5 is the better of these two with the value .32. The rest of the items (n=3) are mediocre and the median is .18.

In quiz 3, item 4 (.15) is the most difficult item of reading section and only 15% of students answered this item correctly. On the other hand, item 3 (.85) is the easiest of all reading items with 85% of students answering it correctly. This exam has the median of .68. In quiz 3, item 4 is the only difficult item and 7 out of 10 items are classified as mediocre. There are also two easy items in these exams (values between .80 and .85).

Item 9 has a negative discrimination value (-.03) and this is the only negative value of reading items in both quizzes. This means this is not a good item in terms of discrimination and it should be either discarded or revised. The best item in terms of discrimination in quiz 3 is item 10 (.29), but this value means even the best item is mediocre (median=.21). Except item 9 which has a negative value, 4 items need revision and the rest of the items (n=5) are classified as mediocre.

When we look at both quizzes, there is only 1 difficult item and we can say that the difficulty of both exams are not very high. Just over half of the items (n=12) are categorized as mediocre and the rest (n=7) are classified as easy. In terms of item discrimination, approximately half of the items (n=9) need revision and there is 1 item which needs to be discarded, so revision seems necessary for better discrimination.

Midterms

For item difficulty and discrimination values and labels for reading items in midterms, refer to Table 23 below.

Table 23

Reading Item Analysis (Midterms)

	Difficulty			Discrimination			Difficulty Label			Discrimination Label		
Items	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Reading 1	.89	.89	.73	.12	.17	.32	E	E	M	NR	NR	G
Reading 2	.98	.56	.75	.10	.40	.28	E	M	M	NR	VG	M
Reading 3	.88	.65	.89	.07	.17	.16	E	M	E	NR	NR	NR
Reading 4	.36	.48	.71	.26	.14	.31	M	M	M	M	NR	G
Reading 5	.56	.63	.70	.25	.22	.31	M	M	M	M	M	G
Reading 6	.33	.80	.95	.13	.36	.14	M	E	E	NR	G	NR
Reading 7	.65	.81	.70	.09	.34	.34	M	E	M	NR	G	G
Reading 8	.78	.43	.88	.21	.24	.17	M	M	E	M	M	NR
Reading 9	.69	.17	.69	.44	.18	.26	M	D	M	VG	NR	M
Reading 10	.88	.69	.96	.24	.42	.17	E	M	E	M	VG	NR
Reading 11	.82	.29	.63	.14	.38	.25	E	D	M	NR	G	M
Reading 12	.92	.63	.80	.19	.24	.10	E	M	E	NR	M	NR
Reading 13	.20	.64	.81	-.02	.21	.30	D	M	E	DIS	M	G
Reading 14	.45	.80	.75	.22	.29	-.01	M	E	M	M	M	DIS
Reading 15	.62	.77	.62	.14	.31	.21	M	M	M	NR	G	M
Reading 16	.45	.81	.87	.36	.26	.07	M	E	E	G	M	NR
Reading 17	.68	.55	.95	.19	.10	.22	M	M	E	NR	NR	M
Reading 18	.45	.80	.64	.20	.18	.25	M	E	M	M	NR	M
Reading 19	.95	.74	.70	.16	.24	.27	E	M	M	NR	M	M
Reading 20	.65	.91	.60	.28	.28	.27	M	E	M	M	M	M
Reading 21	NA	NA	.49	NA	NA	.29	NA	NA	M	NA	NA	M
Reading 22	NA	NA	.71	NA	NA	.32	NA	NA	M	NA	NA	G
Reading 23	NA	NA	.42	NA	NA	.28	NA	NA	M	NA	NA	M
Reading 24	NA	NA	.76	NA	NA	.26	NA	NA	M	NA	NA	M
Reading 25	NA	NA	.78	NA	NA	.13	NA	NA	M	NA	NA	NR

Note. Each item in the first column corresponds to a specific item in each exam. E stands for easy, M stands for mediocre, D stands for difficult, NR stands for needs revision, DIS stands for discard, G stands for good and VG stands for very good. NA stands for not available.

Item difficulty values for reading items in midterm 1 range between .20 and .98. Item 13 is the most difficult item while item 2 is the easiest. Median for item difficulty values is .67. In midterm 1, there is 1 item with the value below .30, so there is 1 difficult item in this exam. Twelve out of 20 items range between .33 and .78. These items are categorized as mediocre. There are 7 easy items which have values .82 or above.

Item 13 has the value of -.02 and this negative value makes it the weakest item of all in terms of discrimination in midterm 1 (median=.19). In contrast to item 13, item 9 (.44) is the most powerful item which is classified as very good. As we have 1 item with negative value, this item should be discarded or revised. We also have 1 good (.36) and 1 very good (.46) item. Half of the items need revision for better discrimination and there are 7 mediocre items whose values range between .20 and .28.

There are items of all item difficulty labels in midterm 2. Two of the items are difficult and between these two, item 9 is the most difficult item (.17) with only 17% of students who answered it correctly. On the contrary, item 20 is the easiest (.91) item of midterm 2 (median=.67). Slightly above half of the items (n=11) are mediocre items (values between .43 and .77) and the rest of the items (n=7) are easy.

We also see all labels (except for discard) in terms of item discrimination in midterm 2 which means there are mediocre, good and very good items as well as those that need revision. While item 17 (.10) is the weakest item, item 10 (.42) is the best item that discriminates between high achievers and low achievers (median=.24). When we look at the distribution of these labels among items, there are 6 items which need revision and 8 out of 20 items are mediocre, 4 of them are good items and 2 of them are very good items.

Unlike midterm 1 and 2, we do not have any difficult item in midterm 3. The most difficult item has a value of .42 (item 23) which makes it mediocre and the easiest item has the value of .96 (item 10) (median=.73). We understand that even the most difficult item was correctly answered by almost half of the students. Most of the items (n=17) are classified as mediocre in this exam and there are also 8 easy items (values between .80 and .96).

Item discrimination values of midterm 3 (median=.26) show that most of the items are either mediocre or good items. Although the overall discrimination level is acceptable, there is 1 negative value of -.01 (item 14) which should be discarded from the exam. The best item of midterm 3 is item 7 (.34). Except for the item with a negative value, 7 items also need revision as their values are below .17. Just under half of the items (n=11) are mediocre and 6 items, whose values range between .30 and .34, are good in terms of item discrimination.

When we look at all the midterms in terms of item difficulty for reading items, only 3 items are classified as difficult and a great majority of items (n=40) are mediocre in terms of difficulty. The rest (n=22) are easy items, which suggests that the difficulty of reading items are not very high. However, as most of the items are mediocre, this is quite good for the general difficulty of midterms. As for item discrimination values, two items which have negative values do not work well for discrimination. Approximately one third of the items (n=23) need revision and 26 items are categorized as mediocre. So, it can be stated that the items may be revised as a whole for better discrimination and the number of good and very good items may be higher.

If we compare the difficulty values and discrimination labels of reading items, out of 29 easy items, 19 of them need revision and 7 of them are mediocre in terms of discrimination. The number of mediocre items (in terms of difficulty) in all exams is 52 and only 12 of them need revision while there are 25 mediocre, 10 good, 3 very good items and 2 items to be discarded. When we look at the ratio of items that need revision, more than half of the easy items need revision while only less than a quarter of mediocre items need revision. Thus, easy items are mostly the ones that need to be revised.

CHAPTER 5: DISCUSSION

Introduction

In this chapter, overview of the study was given and following this, major findings and conclusions about research questions were shared. Major findings and conclusions part was given based on domains. Each domain was evaluated for item-objective agreement first and then evaluation of item difficulty and discrimination was provided with references to literature. After major findings and conclusions part, implications for practice and further research were shared and lastly, limitations of this study were provided.

Overview of the Study

One of the purposes of this study was to examine to what extent tests match with learning objectives. The second purpose of the study was to investigate parameters of items used in the assessment materials. Alignment of test items with learning objectives is of significant importance. However, this is not enough. Unless items have good quality as shown by item parameters such as difficulty and discrimination, alignment would not be meaningful. In the first stage, agreement level between listening, grammar, speaking and reading domains' objectives (both course objectives and course book objectives) and assessment tools (quizzes and midterms) was analyzed separately. In the second stage, item difficulty and item discrimination parameters were calculated using students' responses.

Major Findings and Conclusions

In this part of the chapter, major findings and conclusions based on the objective-item matching and item parameters were provided. First, findings about objective-item matching were shared and each domain was given separately as in Chapter 4. Then, findings and conclusions about item difficulty and discrimination analyses were shared following the same pattern.

Listening Domain

Table 24 below shows the number of items measuring the listening course objectives and course book objectives.

Table 24

Number of Items Measuring Listening Objectives

Course Objectives	*f*	Course Book Objectives	*f*	Number of Items
CO1	75	CBO1	4	85
CO2	35	CBO2	75	
CO3	5	CBO3	0	
CO4	27	CBO4	27	

In all exams throughout one semester, there were a total of 85 listening items in different assessment tools. When we look at the distribution in Table 24 above, we notice that understanding details and specific information objective (CO1 & CBO2) was tested in most of the items (n=75) and other objectives were tested in fewer items. Normally, this is not a sample of a desired distribution in a test which lowers the content validity as well. Pilliner (1968) asserted that the relationship between test items and learning objectives affects the content validity and Heaton (1988) also commented on the distribution of items in a test by claiming that tests should have items that represent the whole course content and learning objectives. Gronlund (1977) suggested using a table of specification in order to avoid this problem. He added that table of specification is quite useful to balance the weight of each objective and content of the course. Based on this view, a table of specification might be especially useful to cover all the objectives.

When we look at the items, all of them align with objectives as shown by 100% agreement between listening items and objectives (both course objectives and course book objectives). When Sireci's domain representation and domain relevance elements are taken into consideration, we can see that all of the aspects of listening domain were measured in the exams and there were no irrelevant items. Cohen and Wollack (2004) also emphasized the importance of the agreement between objectives and items. They suggested that each objective should be tested with one or more items. Based on this agreement level between the tests items and learning objectives, we can say that students were tested on what they were expected to learn which is quite important for teaching and assessment. On the other hand, the weightings of topics (the amount of time spent to teach the topic) in the course book should also be taken into consideration to have a clearer picture about the item distribution.

Almost all COs and CBOs were tested with at least 1 item. This is quite desirable for testing. As Bannister and Rochester (1997) claimed covering all the objectives in a test is crucial to be sure of students' learning. A great majority of items measured students' ability to understand specific details in short and long conversations or lectures and making inferences, but students' ability to interpret speaker's tone and attitude was not measured. CO3 was measured with 5 items in midterm 3 as students listened to a broadcast radio documentary to answer these items. However, as CO3 has two parts, (understanding the broadcast audio and identifying the speaker's tone and attitude) only 1 part (understanding the broadcast audio) of this objective was measured. This suggests that maybe CO3 should be revised and understanding the broadcast audio and identifying the speaker's tone and attitude parts should be different objectives as they are two different skills. That is because students may understand the broadcast audio, but they may not identify the speaker’s tone or attitude or the opposite may happen and although students can identify the speaker’s attitude, they may not understand the whole audio. So, these are two different aspects and they need to be in different objectives as it is impossible to understand which part of the objective is met. Albritton and Stacks (2016) shared similar opinions. They claimed that, if necessary, objectives can be edited to focus more on key information or key skill. Moreover, understanding the speaker’s tone and attitude items can be asked in different contexts such as long talks or lectures as well.

If we are to reach a conclusion about the listening items, it is quite clear that listening items were mostly based on understanding the details. Approximately a quarter of items also measured students’ inference skills. As a great number of items measured understating details, less than 5% of items focused on understanding the main idea of talks or long conversations. Surprisingly, none of the items was about speaker’s tone and attitude although it is one of the objectives. This might hinder effective teaching as Mogapi (2016) posited that objectives which are not measured have an indirect negative effect on teaching because nothing is known about these objectives.

However, this distribution seems quite fair considering TOEFL ITP or other high-stake exams. The institution is preparing its exams mostly in the same pattern with TOEFL ITP and in TOEFL, most of the items measure understanding specific details and information. So, considering TOEFL and the nature of listening exams, it is quite expected that the number of items measuring the specific details objective is quite high (n=75). This is an important skill in high stake exams, so it was measured almost in all items. What is more, as short conversations do not have a main idea and each lecture or talk only has 1 main idea, the number of items measuring main ideas can only be limited in number. So, in TOEFL ITP, there are very few items measuring this skill (in some talks and long conversations there is no item), and similar to this, items in the institutions' exams did not measure understanding the main idea objective much.

On the other hand, as understanding speaker's tone and attitude objective was never tested, without doubt, items measuring this skill should also be added in the exams because if this is one of the aims, then the institution cannot be sure whether the students master this skill or not. Furthermore, it is an important skill for TOEFL since this objective is tested. As Gronlund (1977) suggested the first step to assessment process is to identify the objectives that should be tested. If this is done, objectives will not be skipped and each objective can be measured. Furthermore, these objectives that were not measured in the exams create content validity problems as all the content was not addressed with items.

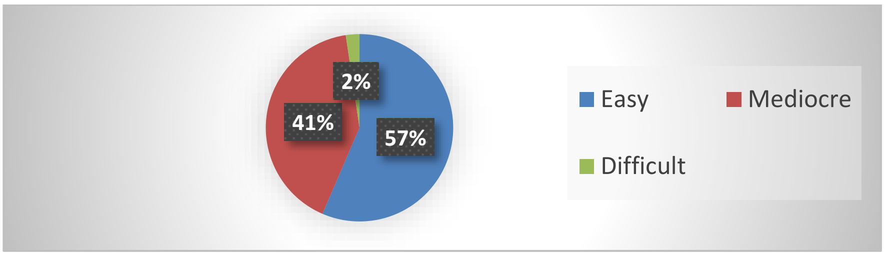
Table 5 and Table 6 in Chapter 4 show that some items measured 2 objectives (COs) at the same time both in quizzes (n=12) and midterms (n=42). This is also the case for CBOs (quizzes=5 & midterms=19). At first sight, this is not something desired in terms of testing. As Hambleton and Eignor (1979) suggested each item should ask for single information, in other words, should test 1 objective at a time. Similarly, Brown and Abeywickrama (2010) also stated that each test item should only match with 1 objective. On the other hand, the reason for this situation may be because understanding specific details and information skill is mostly tested with understanding extended speech and lectures, understanding short conversations or making inference objectives. This is the case both for COs and CBOs. I believe that, as I mentioned above, although focusing on specific objectives rather than all of them is a problem according to literature, this is not a problem as these objectives are highly inter-related. To clarify, students listened either short conversations or long conversations and talks in the exam. When they listened to a long conversation and talk, most items measured understanding specific details skill. So, such items match both of the objectives. That is because students had to understand the long conversation and talk to answer specific details items correctly. Likewise, in some items, students had to understand short and long conversations or talks and needed to make an inference. Thus, such kind of items also measured 2 objectives at the same time.

Although measuring more than 1 objective is not very problematic for listening items due to the abovementioned reasons, some objectives can still be revised so that items might match only 1 objective. Richards (2002) supported this view by stating that learning objectives can be checked and can also be narrowed down into smaller and observable objectives. Besides more uniform distribution, merging objectives might be a solution for this. To illustrate, CO2 (understanding extended speech, long conversations and lectures and follow even complex lines of argument provided the topic is reasonably familiar) can be revised because when we look at this objective, the only aim is to understand. However, this objective might be rewritten as "understand specific details by following extended speech, long conversations and lectures and follow even complex lines of argument provided the topic is reasonably familiar". So, most of the items would match only with this objective instead of 2 objectives.

Figure 7 shows listening items' difficulty level distribution.

Figure 7

Distribution of Difficulty Levels for Listening Items



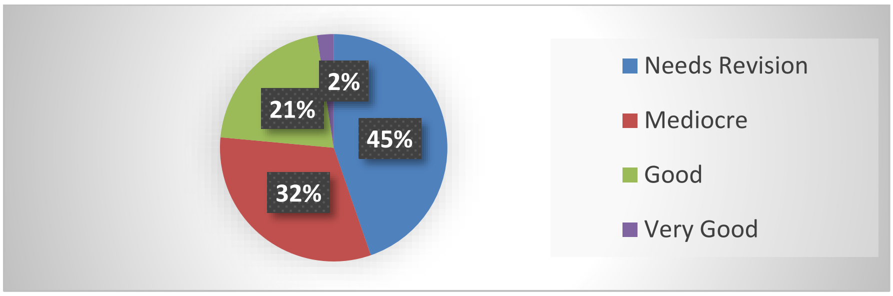
When we look at the item difficulty of listening items, items mostly consist of easy and mediocre items. Having such a high number of mediocre items is quite good; on the other hand, number of easy items might be reduced and difficult items might be increased. Ebel and Frisbie (1991) shared their views on ideal item difficulty. They stated that in order for good assessment, items should neither be too easy nor difficult. Item which are mediocre in difficulty are the best as they discriminate students better. Henning (1987) shared similar views and claimed that if an exam is very easy or very difficult, it may lose the ability of discriminating students which causes decrease in reliability. Median for item difficulty in quizzes is .90 and for midterms .80. That is, most of the items in quizzes were easy, few of them were mediocre and there was just 1 difficult item. We can conclude that midterm items were more difficult than quiz items. This might be because midterms covered all topics covered throughout the semester, but quizzes covered the topics of 2 or 3 weeks' period. Thus, students could answer the quiz items more easily.

If we compare the difficulty levels of items which measured 1 objective and 2 objectives, items measuring 2 objectives have the median of .85 while items measuring 1 objective have .89. These values show that both group of items have approximately the same difficulty level and we can conclude that whether the item measures 1 or 2 objectives does not affect its difficulty much in listening items.

Figure 8 shows the distribution of discrimination labels for listening items.

Figure 8

Distribution of Discrimination Labels for Listening Items



If we look at the overall values, considering Ebel and Frisbie's (1991) suggestions, items need to be revised to be better in terms of item discrimination as approximately half of the items need revision according to Figure 8. Similarly, there should be a higher number of good and very good items. In terms of item discrimination, median for quizzes is .17 and for midterms it is .24. So, discriminating power of midterms was higher than quizzes'. As we can understand from the values, most of the items which have values less than .29 either need revision or are mediocre in both exams. Median for items measuring 1 objective is .19 and for items measuring 2 objectives is .24. Thus, we can clearly see that although there is not much difference, items measuring 2 objectives have higher level of discrimination.

Grammar Domain

In Table 25 below, number of items measuring grammar objectives is shown.

Table 25

Number of Items Measuring Grammar Objectives

Course Objectives	*f*	Course Book Objectives	*f*	Number of Items
CO1	28	CBO1	5	100
CO2	0	CBO2	8	
CO3	0	CBO3	30	
CO4	8	CBO4	10	
CO5	0	CBO5	3	

Table 25 (cont'd)

Number of Items Measuring Grammar Objectives

CO6	0	CBO6	7
CO7	55	CBO7	11
CO8	8	CBO8	13
CO9	19	CBO9	4
		CBO10	4
		CBO11	4
		CBO12	1
		CBO13	8

To summarize Table 25 above, CO7 (producing well-structure sentences) was tested in 55 items which means slightly above 50% of items measured just this objective. As grammar mainly focuses on structure, it is quite acceptable to test this objective in many items. Similar to Heaton's (1988) view, structure items are quite appropriate to be tested in grammar section. CO1 (using proper language for different aims) was also tested in 28 items. So, 83 out of 100 items measured these 2 objectives only. Having such high frequency for just 2 objectives seems like a problem, but actually it is not for grammar domain. They are the basics of grammar and considering the nature of grammar tests, it is quite expected that they were tested in so many items. Even in high stake exams such as TOEFL ITP, all grammar items test students' ability to use proper language and their structure knowledge. However, the items should cover more objectives (all of the objectives if possible) to be sure that students can do all the tasks stated in the objectives. To exemplify, students' ability to form structures for complaints and directions were not tested, but they can also be tested even in grammar context and they should be integrated into exams.

Among 100 items, only 1 item did not match with any of the course objectives and 7 items did not match with course book objectives. We can say that agreement level of grammar items is quite high (99% for course objectives and 93%

for course book objectives). Although there are few items which do not match with objectives (especially CBOs) and it seems quite problematic and not appropriate for measurement and assessment, there is an explanation for this situation. For some grammar topics which were not included in the course book, students were given extra handouts and they studied the grammar topic in their classes with the help of these handouts. So, these items measured the skills which were taught in these handouts. So, in reality, although there is no match with the course book objectives for 7 items, there is a match with handouts, so the agreement level is even higher (considering that these 7 items match with the objective handouts, we can say that the agreement level between items and course book objectives is 100%).

When we look at the distribution of course objectives, among 9 objectives, 4 of them (objectives 2, 3, 5 and 6) were never tested. This means students were not tested for their ability to understand and produce simple tasks, form structures for directions, deal with routine situations and form structure for complaints. The reason for this situation is that COs were mostly written based on CEFR and they were very general statements. What’s more, they were written in a very general sense rather than focusing solely on grammar. Actually, grammar points that students learned help them to understand simple tasks, deal with routine situations or form structures for directions. So, although they were not tested directly in the exams, if students master grammar points, they can also do the abovementioned tasks. The reason that they were not tested directly is that it might be difficult to test these objectives in the grammar section of the exams because in grammar section of almost all kinds of exams, grammar structures and forms are the main focus. Heaton (1988) supported this view and asserted that certain skills might be considered more appropriate to be tested with specific type of items and other skills might be ignored. That’s why there are some missing objectives in the exams. Furthermore, in terms of Sireci's (1998) content validity elements, all the objectives of grammar domain were not represented in the exams, but almost all items were relevant to grammar domain. In order to be sure of a desired distribution among the objectives, as Gronlund (1977) suggested table of specification might be of great importance. Thanks to table of specification, it can be easily noticed that there are some missing objectives in the exam. On the other hand, the institution may still revise the grammar objectives because as they were not tested, then maybe they are not the main objectives for this class and the institution may consider removing these objectives or editing them.

In terms of course book objectives, the distribution seems a lot better. All of the objectives of the course book were measured in the exams. Although some objectives were tested with much more items, it is also because these topics are much more appropriate to be tested in grammar sections. Furthermore, another reason of this distribution might be students studied some grammar topics at the beginning of the semester and some towards the end. So, if they had studied the topic at the beginning of the semester, they could be tested on this topic throughout the semester in different exams and this means more items tested these objectives. On the other hand, if they learned the grammar topic at the end of the semester, they could only be tested in midterm 3 or maybe in the last 2 exams. This might be the reason why some objectives were not tested in midterm 1 or midterm 2. This lowers the number of items measuring some of the objectives as well.

Figure 9 & Figure 10 show the distribution of difficulty levels and distribution of discrimination labels for grammar items.

Figure 9

Distribution of Difficulty Levels for Grammar Items

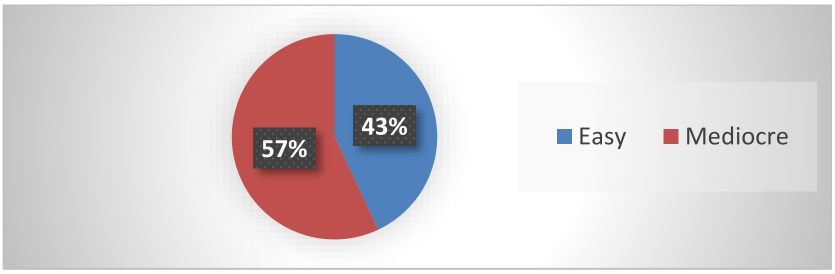
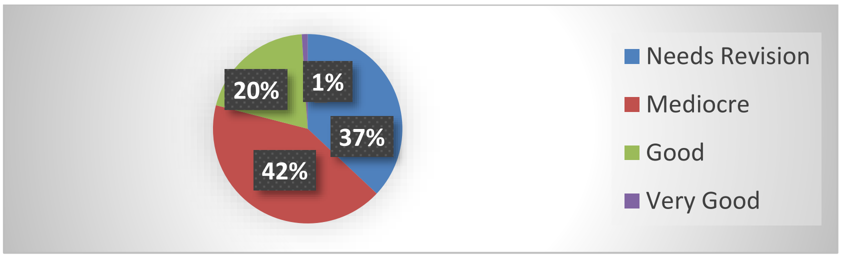


Figure 10

Distribution of Discrimination Labels for Grammar Items



Item difficulty level of grammar items show that items are a little easy and there should be difficult items in the exams as well (there is no difficult item in any exams). The number of mediocre items (n=57) seems quite fair. Thus, adding some difficult items can be better for better assessment. Item discrimination labels and values suggest that some improvements can be made for better discriminating power. Approximately one third of items need revision, so this number should be reduced by revising the items. Only one fifth of the items are either good or very good, so this number should also be increased.

When we compare the difficulty values of quizzes and midterms, quizzes have the median of .73 and midterms have .76. So, we can clearly say that difficulty of quizzes and midterms are almost the same which is quite positive for the assessment. When we look at the discrimination values, quizzes have the median of .18 and midterms have .26. By looking at the values, it is quite clear that midterm items discriminated between students much better when compared to quizzes. This might be because quizzes cover topics for every 2 or 3 weeks, but midterms cover topics from the beginning of the semester and some students might have just focused on or memorized the rules and got good grades in the quizzes, but they might not have done well in midterms. That's why discriminating power of midterms is higher although the difficulty levels are almost the same.

If we look at the items measuring 1 objective (n=80) (median=.76) and 2 objectives (n=19) (median=.79), difficulty values are almost the same for both group, so we can say that whether the item measures 1 objective or 2 does not have an effect item difficulty. The discrimination values of items measuring 1 objective (median=.24) and 2 objectives (median=.20) show that items measuring 1 objective are slightly better in terms of discriminating power. According to literature, items should measure only 1 objective, so considering this view, grammar items mostly obey this rule. Cohen and Wollack (2004) expressed their ideas on this issue by saying that an item should measure just one objective as much as possible.

Speaking Domain

Table 26 shows the number of items measuring the speaking objectives.

Table 26

Number of Items Measuring Speaking Objectives

Course Objectives	*f*	Course Book Objectives	*f*	Number of Items
CO1	14	CBO1	1	15
CO2	0	CBO2	0	
CO3	5	CBO3	2	
CO4	3	CBO4	4	
CO5	0	CBO5	14	
CO6	2			
CO7	4			
CO8	0			
CO9	0			

When we look at the agreement level between speaking items and objectives, all items match with course objectives and except for 1 item, all items also match with course book objectives. There are a total of 9 course objectives for speaking and out of 9 objectives, 5 of them were measured, but 4 of the objectives (objectives 2, 5, 8 and 9) were not measured in speaking quiz. We can conclude that students were mostly measured for their ability to state ideas and preferences, take part in a conversation, describe personality, talk about similarities and differences and state ideas on contradictory issues. So, although they are the objectives of this course, students' ability to make a presentation, make corrections, perform role plays and summarize were not measured. For course book objectives, only 1 objective (explaining cause and effect) was not measured directly. However, students might have used cause and effect relation when they talked about their ideas or speculating about future. So, this objective might have been measured indirectly during the exam. When we consider Sireci's (1998) four elements of content validity, almost all items were relevant to speaking domain; however, all objectives of speaking domain were not represented.

As I stated above, some objectives were not tested, but they are listed in the course objectives. If they are not tested, the institution cannot know whether the students can do these tasks or not. As Kirkpatrick and Kirkpatrick (2006) stated if the objectives are not covered, their attainment levels cannot be observed. Geerts et al. (2018) supported the view that at the end of each course, institutions should measure whether the objectives are met or not. Similarly, Alonso et al. (2008) claimed that objectives need to be measured to observe whether they are met or not, and thanks to this evaluation, necessary changes can be made on courses. So, if they are not the institutions' primary objectives, then objectives might be revised or items covering these objectives can be developed and added. If they are main objectives of the institution, they can also be integrated in the assessment process somehow.

As some objectives regarding making presentation, summarizing or performing role plays were not never measured in speaking exam, we can come to a conclusion that the institution uses the same type of assessment tools or items to assess students' speaking skills and with more variety, these objectives can be easily addressed as well. Luoma (2004) suggested some alternatives for speaking assessment which are pair, group and communication-oriented tasks and added that they are also not time consuming as one to one assessment. These two kinds of tasks can also be integrated in the speaking assessment for the missing objectives. Thanks to group tasks or communication-oriented tasks, students can be asked to make corrections to each other during the assessment or perform role plays with role cards. Brown (2003) also suggested some alternatives for speaking assessments which are retelling a story and read aloud tasks. With these tasks, students' summarizing skill, which was not tested in the speaking exam, can be tested. Students might be asked to read aloud a text first and then summarize it in a minute.

Similarly, they may listen a story and then they might be asked to retell a story in their own words by summarizing it. These methods might be good options to test students' summarizing skills which is among the objectives of the course. Among other alternatives suggested are Coulson's (2005) "team talking" tasks which includes use of communication strategies and Skehan and Foster's (2001) problem solving and debate tasks. These tasks can also be good alternatives for role plays and making corrections or even making presentations.

Reading Domain

In Table 27 below, number of items measuring reading objectives are presented.

Table 27

Number of Items Measuring Reading Objectives

Course Objectives	*f*	Course Book Objectives	*f*	Number of Items
CO1	8	CBO1	7	85
CO2	0	CBO2	0	
CO3	7	CBO3	12	
CO4	43	CBO4	43	
CO5	14	CBO5	0	
CO6	6	CBO6	14	
CO7	0	CBO7	6	
		CBO8	2	
		CBO9	0	
		CBO10	8	
		CBO11	2	
		CBO12	0	

Table 27 above shows that two-thirds of items (n=43) measured students' scanning skills. And the rest of the items measured skimming skills, guessing the meaning or sentence paragraph relationship. This distribution might again seem problematic as there is too much focus on 1 objective; however, I think it is not very problematic since most of the reading items in other exams also measure students' scanning skills which require students to find specific details and information in the text. Scanning is a very important skill in reading and mastering it helps students to be more proficient in English (Wahyuni et al., 2014). When we think about reading items in high stakes examinations such as TOEFL or IELTS, the distribution is much like the same. Brown (2003) mentioned some of the TOEFL specifications for reading and added that they are quite in accordance with effective reading skills. He also suggested using the same format and type of items for reading assessment.

Furthermore, some objectives cannot be tested with a lot of items. For example, there are 10 items for each reading text and although main idea is the most significant component as it gives the writer's purpose (Pierce & Kinsell, 2008), each text has just 1 main idea which means there can only be 1 item regarding the skill of understanding the main idea. So, the number of items measuring this skill cannot be more than the number of texts used in exams (if in each text there is a main idea question). That's why the frequency of CO3 is not very high. Similarly, CO6 was also measured in 6 items because the objective is about understanding the organization of the texts and CO6 also cannot be measured in many items as each text has only 1 type of organization.

When we look at the objectives of the reading course, they are also quite consistent with the item types of high stake exams. Brown (2003) stated reading item types used in TOEFL exams which are understanding of main ideas and factual information, making inferences, vocabulary and pronoun referents. We can clearly notice that these are among the reading course objectives (except for pronoun referents) and all of them were tested in quizzes and midterms. Out of 7 course objectives, 2 of them (objective 2 and 7) were not tested in any exam. So, students' ability to use dictionary (objective 2) and differentiate between facts and opinion (objective 7) were not tested. Especially the 2nd objective is not appropriate to be tested in exams as students are not allowed to use dictionary in any exam. So, it is quite expected that objective 2 was never tested. However, items regarding the 7th objective can be added to determine whether the students can do the differentiating or not. Differentiating between facts and opinions is an important skill for students to become more effective and critical readers (Albeckay, 2014).

When we look at the course book objectives, out of 12, 4 objectives (objectives 2, 5, 9 and 12) were never tested. This means students' ability to identify the purpose of the text, find synonyms, summarize and distinguishing between fact and theories (similar to course objective 7) were not tested. Among these objectives, finding synonyms can be integrated into vocabulary items and instead of asking for the meaning or definition of words, synonyms of words can be asked to students to cover this objective. Purpose of the text objective can also be measured easily, but like main idea, the number of items can be limited as there is just one purpose or organization for each text.

In terms of agreement level of items and objectives, 71 items match with course objectives and 79 items match with course book objectives. So, 14 items do not match with any of the COs and 6 items do not match with any of the course book objectives. So, when compared to other domains like listening, grammar or speaking, the agreement level is relatively low for reading domain. Moreover, a great majority of items measure only 1 course objective (n=59) and course book objective (n=50). We can see that most of the reading items stick to the rule which is each item should measure only one objective.

As Grabe (2009) suggested students should be tested on a variety of skills while assessing reading, so objectives which were not tested in the exams (except for using dictionary) should also be integrated. While testing the reading skills, only multiple-choice items were used in quizzes and midterms and variety in item types may be another choice for the institution. Kusiak (2002) suggested open-ended items to test reading skills and said that this type of items can be used to test students' skimming and scanning skills. Similarly, Carrell et al. (1989) emphasized the importance of short answer items claiming that they have better discriminating power. Furthermore, objectives should also be revised and some objectives may be added or some of the can be modified. For example, a great majority of items which do not match with course objectives are pronoun reference items. Pronoun reference items were tested almost in all exams and it is also a type of question which is frequently tested in high stake exams such as TOEFL. So, understanding pronoun references must be among the objectives and it can be added to course objectives. Furthermore, items covering objectives which were not covered in any exam can be added.

Figure 11 and Figure 12 below show distribution of reading items' difficulty levels and discrimination labels.

Figure 11

Distribution of Difficulty Levels for Reading Items

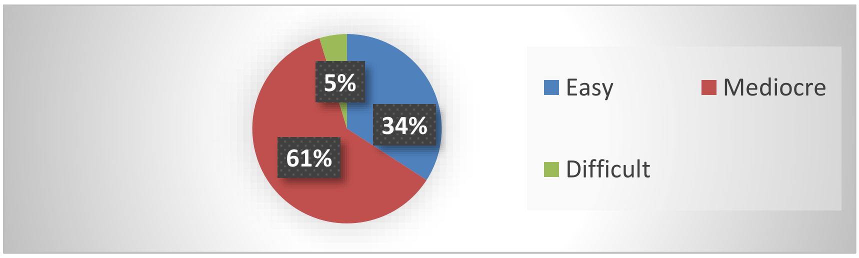
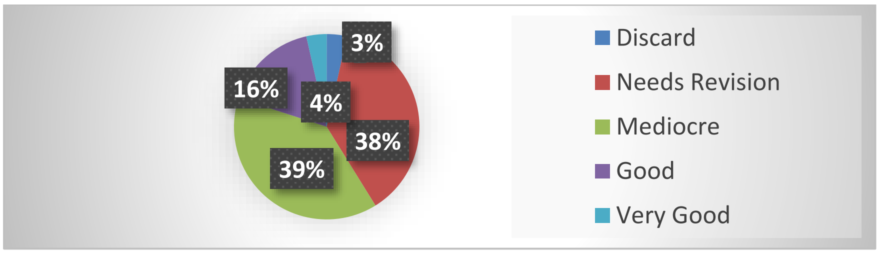


Figure 12

Distribution of Discrimination Labels for Reading Items



Such a high number of mediocre items is quite good for an assessment tool in terms of difficulty, but still there is room for improvement. Number of difficult items might be higher while easy items may be lower. For the item discrimination, this domain is the only domain that has items to be discarded and in terms of discrimination, the number of good and very good items should be higher.

When we compare reading items in quizzes and midterms in terms of item difficulty, quizzes (median=.74) and midterms (median=.70) have almost the same difficulty level. In terms of item discrimination, quizzes (median .20) and midterms (median .24) have similar discriminating power. So, similar item parameters show that reading items in quizzes and midterms are quite similar.

When we look at the difficulty values of items which measure only 1 objective, 2 objectives and which do not match any objectives, median for each category is .71 (mediocre). So, we can say that whether the item measures 1 or 2 objectives or matches with learning objectives or not do not affect the item difficulty. However, when we look at the item discrimination values of these 3 groups, items measuring 1 objective have the median of .24 (mediocre), items measuring 2 objectives .31 (good) and items which do not match any of the objectives .17 (need revision). These values suggest that items measuring 2 objectives have the best discriminating power while items which do not match any of the objectives have the lowest discriminating power. Thus, item-objective agreement has an effect on item discrimination.

Summary of Discussion

The purpose of this section is to summarize the main points discussed in Chapter 5. Firstly, distribution of objectives can be taken into consideration. In some skills, some of the objectives were measured a lot more than others. To some extent, it is acceptable due to the abovementioned reasons; however, the distributions still need to be revised because there is still room for the improvement.

Some objectives of certain skills were not measured in any exam. This is not appropriate for testing. For some skills like grammar, some of the objectives are not suitable to be tested in the grammar part as the main focus is on structure and the form. Of course there are different methods to test grammar, but the institution uses the same item types with TOEFL ITP, that’s why it is not appropriate to test some objectives. However, they might be tested in few items. For other skills, it is much easier to test all objectives.

Some objectives might need revision. To illustrate, some of the objectives have two different aspects to be tested. Some items might measure one part of the objective but not the whole objective. So whether the objective is achieved or not cannot be understood. This is problematic and these objectives need to be revised. Also, some of the objectives that were not tested might not be among the main objectives of the class. These objectives need to be revised as well.

Some of the items need to be better in quality. As stated in Chapter 4, there are some items that need revision in terms of item difficulty. Also, the number of difficult items should be increased in general as there are very few and even no item. Moreover, some items also need revision for item discrimination. There might be many reasons why items do not discriminate well. It might result from the item stem, distractors or some other reasons. Furthermore, as the number of items which were categorized as easy can be considered high, this might also lead to low discriminating power. To sum up, some items need revision for difficulty and discrimination.

Furthermore, some items measure more than 1 objective which seems problematic in terms of testing. However, for some items this is not a problem because some objectives are quite related to each other. This problem can be solved not only revising the items but also revising the objectives.

Lastly, for some skills, alternative question types or assessment tools can be used to address the objectives because it is very difficult to address all the objectives with only multiple-choice items in most skills and interview type items in speaking exams. Variety of assessment tools and items can help the institution to address all the objectives.

Implications for Practice

- Curriculum unit members may be provided with a training on how to write/evaluate objectives,
- Testing unit members may be provided training on how to develop items considering different aspects of item development such as item stem, distractors, balancing item difficulty, addressing the objectives and editing items based on item parameters,
- Table of specification may be prepared prior to item development for each exam,
- In each exam, pilot study on some items can be conducted. Extra items, which are not taken into consideration while calculating students' grades, can be added into each exam. If item parameters are good for these items, they may be used in the exams which will be given in the following years,
- Instructors at preparatory school should also be reminded about the learning objectives and the importance of them to be covered in classes and exams.

Implications for Further Research

Distractor analysis is highly important to improve item quality. Haladyna (2004) suggested that the discriminating power of an item is highly related to distractors. On the other hand, writing reasonable distractors is not an easy task. Considering this view, distractor analysis can also be conducted to improve item quality. In this study, it was not conducted as the data were not available. (The only available data were “1” for correct answers of students and “0” for wrong answers. So, the distribution on options was not available). Furthermore, besides distractor analysis, other reasons that affect item discrimination can also be analyzed by examining the items that have low discriminating power.

Another implication might be conducting item stem analysis. Item stem might affect not only item difficulty but also item discrimination. The grammatical structure or vocabulary used in item stem should be appropriate for students’ level; otherwise, students might not give correct answer to a question just because they do not understand the item stem or they do not know some words in the item stem. To illustrate, if we are testing students’ grammar and if the students cannot answer the question just because they do not know some words and they cannot understand the item stem, we are not testing grammar here, instead we are testing students’ vocabulary. So, item stem is also very important for item quality.

Furthermore, as the number of difficult items is quite low in the exams in general, items can be checked whether some of them give clues for other items. In other words, students might answer some questions with the help of other items or item stems. For example, if an item is measuring students’ ability to use conditional type 2 and if in another item stem, conditional type 2 is used, students can answer the question easily even if they do not know the answer. So, items might be checked in this respect.

In this study, matching method was used to analyze the alignment level of learning objectives and test items. Besides matching method, rating method can also be used to analyze this alignment level. For each item, a numerical score can be assigned to indicate to what extent the items match with the objectives.

Moreover, for the item analysis part, item analyses based on IRT can be conducted. By this way, significant information can be obtained from a small pilot study group and item quality can be generalized to larger groups of students with different ability levels prior to real testing of the items.

Although writing is one of the main skills of language teaching, it is not analyzed in this study since the students are expected to write a paragraph or essay in the exams. However, if there are other assessment tools for writing in an institution, writing part can also be analyzed.

Limitations

- The data were collected only from 1 semester of the academic year,
- Only quizzes and midterms were analyzed and alternative assessment tools were not analyzed (portfolios),
- Only B level exams and objectives were analyzed (there are A and C levels),
- Only Classical Test Theory was used while conducting item parameters.

REFERENCES

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. *Educational and Psychological Measurement, 42*(3), 803-806.

Albeckay, E. M. (2014). Developing reading skills through critical reading programme amongst undergraduate EFL students in Libya. *Procedia - Social and Behavioral Sciences*, *123*, 175-181. https://doi.org/10.1016/j.sbspro.2014.01.1412

Albritton, S., & Stacks, J. (2016). Implementing a project-based learning model in a pre-service leadership program. *NCPEA International Journal of Educational Leadership Preparation, 11*(1). ISSN: 2155-9635.

Alderson, J. C., Brunfaut, T., & Harding, L. (2017). Bridging assessment and learning: A view from second and foreign language assessment. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 379–387. https://doi.org/10.1080/0969594X.2017.1331201

Alimi, M. M., & Ellece, S. (2003). Course design and testing in an English programme. *Language, Culture and Curriculum*, *16*(2), 244–252. https://doi.org/10.1080/07908310308666672

Alonso, F., Lopez, G., Manrique, D., & Soriano, F. J. (2008). Instructional and technological design of e-learning courses. *New Educational Technology,* (pp. 127-148). Nova Science Publishers.

Alonso, F., López, G., Manrique, D., & Viñes, J. M. (2008). Learning objects, learning objectives and learning design. *Innovations in Education and Teaching International*, *45*(4), 389–400. https://doi.org/10.1080/14703290802377265

Ames, C. (1992). Classrooms: Goals, structures and student motivation. *Journal of Educational Psychology*, *84*(3), 261-271.

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*(2), 153–170. https://doi.org/10.1080/08957340701301272

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

Bannister, M., & Rochester, M. (1997). Performance measures for NSW TAFE libraries: What can we learn from the literature? *Australian Academic & Research Libraries*, *28*(4), 281-296. https://doi.org/10.1080/00048623.1997.10755026

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-75.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry*, *34*(4), 310–316. https://doi.org/10.1176/appi.ap.34.4.310

Brassil, C. E., & Couch, B. A., (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison. *International Journal of STEM Education (6)*16. https://doi.org/10.1186/s40594-019-0169-0

Brown, A. (2012). Ethics in language testing and assessment. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 113-121). Cambridge University Press.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices*. Pearson Longman.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment, principles and classroom practices* (2nd ed.). Pearson Education.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653. https://doi.org/10.2307/3587999

Bygate, M. (2009). Teaching and testing speaking. In M.H. Long & C.J. Doughty (Eds.), *The handbook of language teaching* (pp. 405-440). Blackwell publishing.

Carrell, P., Pharis, B. and Liberio, J. (1989). Metacognitive strategy training for ESL reading. *TESOL Quarterly*, *23*, 647-678.

Carroll, B. J. (1980). *Testing communicative performance.* Oxford University Press.

CEFR, Council of Europe. (n.d.). Retrieved October, 21, 2020, from https://www.coe.int/en/web/common-european-framework-reference-languages

Chiswick, B. R., & Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics*, *13*(2), 246–288. https://doi.org/10.1086/298374

Cohen, A. S., & Wollack, J. A. (2004). *Handbook on test development: Helpful tips for creating reliable and valid classroom tests.* University of Wisconsin-Madison.

Colman, A. M. (2008). *A dictionary of psychology* (3rd ed.). Oxford University Press.

Combs, K. L., Gibson, S. K., Hays, J. M., Saly, J., & Wendt, J. T. (2008). Enhancing curriculum and delivery: Linking assessment to learning objectives.

Assessment & Evaluation in Higher Education, *33*(1), 87–102. https://doi.org/10.1080/02602930601122985

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. The University of Michigan Press.

Coulson, D. (2005). Collaborative tasks in cross-cultural communication. In C. Edwards & J. Willis (Eds.), *Teaching exploring tasks in English language teaching.* Palgrave Macmillan.

Cox, R. C., & Graham, G. T. (1966). The development of a sequentially scaled achievement test. *Journal of Educational Measurement*, *3*(2), 147–150. https://doi.org/10.1111/j.1745-3984.1966.tb00871.x

D'Agostino, J. V., Welsh, M. E., Cimetta, A. D., Falco, L. D., Smith, S., VanWinkle, W. H., & Powers, S. J. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education*, *21*(1), 1–21. https://doi.org/10.1080/08957340701580728

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, *43*(3), 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Fan, J. & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument based validation framework. *Frontiers in Psychology*, *11*, 1-14. https://doi.org/10.3389/fpsyg.2020.00330

Frary, R. B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research and Evaluation*, *(4)*11, ISSN:1531-7714

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge Taylor & Francis Group.

Geerts, W. M., Steenbeek, H. W., & Van Geert, P. L. C. (2018). Assessing situated knowledge. *International Journal of Education and Practice 6*(3), 134-146. https://doi.org10.18488/journal.61.2018.63.134.146

Grabe, W. (2009). Teaching and testing reading. In M.H. Long & C.J. Doughty (Eds.), *The handbook of language teaching* (pp. 441-456). Blackwell publishing.

Graff, M. (2003). Cognitive style and attitudes towards using online learning and assessment methods. *Electronic Journal of e-Learning 1*(1), 21-28.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505. https://doi.org/10.2307/3586500

Gronlund, N. E. (1977). *Constructing achievement tests* (2nd ed.). Prentice-Hall Inc.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). The Johns Hopkins University Press.

Hambleton, R. K., & Eignor, D. R. (1979). A practitioner's guide to criterion-referenced test development, validation and test score usage. *Laboratory of Psychometric and Evaluation Research Report No. 70.*, ED 249 269.

Hartzell, M. S. (1984, April 23-27). *Checking for curriculum/test overlap: Two methods discussed* (Paper presentation). American Education Research Association 68th Annual Meeting, New Orleans, LA, United States.

Heaton, J. B. (1988). *Writing English language tests.* ELBS Longman.

Henning, G. (1987). *A guide to language testing*. Newbury House Publishers.

Higgs, T. V. (1985). Teaching grammar for proficiency. *Foreign Language Annals, 18*(4), 289-296. https://doi.org/10.1111/j.1944-9720.1985.tb01806.x

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly*, *40*(1), 109-131. https://doi.org/10.2307/40264513

Hopkins, C. D., & Antes, R. L. (1990). *Classroom management and evaluation* (3rd ed.). Pencock Publishing.

Ishihara, N. (2009). Teacher-based assessment for foreign language pragmatics. *TESOL Quarterly*, *43*(3), 445–470. https://doi.org/10.1002/j.1545-7249.2009.tb00244.x

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing*, *30*(3), 345–362. https://doi.org/10.1177/0265532213480334

Johnston, B., & Goettsch, K. (2000). In search of the knowledge base of language teaching: Explanations by experienced teachers. *Canadian Modern Language Review*, *56*(3), 437–468. https://doi.org/10.3138/cmlr.56.3.437

Kehoe, J. (1995). Writing multiple-choice items. *Practical Assessment, Research & Evaluation*, *4*(9). Retrieved from https://pareonline.net/getvn.asp?v=4&n=9

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels*. Berrett-Koehler.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage Publications.

Kusiak, M. (2002). What we test when we test foreign language reading skills. In E. Manczak-Wohlfeld (Ed.), *Proceedings of the tenth annual conference of the Polish Association for the study of English: PASE papers in linguistics, translation and TEFL methodology* (pp. 213-220). Jagiellonian University Press.

La Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments.* Council of Chief State School Officers.

Larsen-Freeman, D. (2009). Teaching and testing grammar. In M.H. Long & C.J. Doughty (Eds.), *The handbook of language teaching* (pp. 518-542). Blackwell publishing.

Lawson, K. (2016). *The trainer's handbook*. John Wiley & Sons Inc.

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing*, *28*(3), 367–382. https://doi.org/10.1177/0265532211404188

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79–88. https://doi.org/10.1016/j.jeap.2011.03.001

Ma, K. (2013). Improving EFL graduate students' proficiency in writing through an online automated essay assessing system. *English Language Teaching*, *6*(7), 158-167. https://doi.org/10.5539/elt.v6n7p158

Magno, C. & Ouano, J. (2010). *Designing written assessment for student learning*. Phoenix Publishing House, Inc.

Marzano, R. J. (2006). *Classroom assessment & grading that work.* Association for Supervision and Curriculum Development.

Mayring, P. (2004). Qualitative content analysis. In U. Flick, E. V. Kardorff & I. Steinke (Eds.), *A companion to qualitative research* (pp. 159-176). Sage Publications.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126,* 1-51. https://doi.org/10.1037/rev0000126

McDaniel, M.A., Anderson, J.L., Derbish, M.H. & Morisette, N. (2007). Testing the testing effect in classroom. *European Journal of Cognitive Psychology 19*(4-5), 494-513. https://doi.org/10.1080/09541440701326154

McDermott, K. B., Agarwal, P. K., D'Antonio, L. D., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*(1), 3–21. https://doi.org/10.1037/xap0000004

McDonald, M. E. (2007). *The nurse educator's guide to assessing learning outcomes*. Jones and Bartlett.

McKeown, R. (2001). *Into the classroom. A practical guide for starting student teaching.* The University of Tennessee Press.

McMillan, J. H. (2000). *Basic assessment concepts for teachers and administrators.* Corwin Press.

McNamara, T. (2004). *Language testing*. Oxford University Press.

Mead, N. A., & Rubin, D. L. (1985). Assessing speaking and listening skills. *Eric digest.* Retrieved from ERIC database (ED 263 626).

Miller, M. D., Linn, R. L., & Grounlund, N. E. (2009). *Measurement and assessment in teaching*. Pearson International.

Mogapi, M. (2016). Examinations wash back effects: Challenges to the criterion referenced assessment model. *Journal of Education and e-Learning Research*, *3*(3), 78-86.

Moses T. (2017). A review of developments and applications in item analysis. In Bennett R. & Von Davier M. (Eds.) *Advancing human assessment. Methodology of educational measurement and assessment.* Springer, Cham. https://doi.org/10.1007/978-3-319-58689-2_2

Neuendorf, K. A. (2016). *The content analysis guidebook* (2nd ed.). Sage Publications.

Nitko, A. J. (2004). *Educational assessment of students* (2nd ed.). Merrill.

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(2), 228–249. https://doi.org/10.1017/S0261444811000206

Nozadze, A. (2013). How to make the assessment of grammar skills more efficient? *Journal of Education*, *2*, 25-29.

Palmer, A. (1991). The role of language testing in language program evaluation. In S. Anivan (Ed.). *Issues in language programme evaluation in the 1990s* (pp. 1-15). Seameo Regional Language Center.

Perrin, K. M., & Mayhew, D. (2000). The reality of designing and implementing an internet-based course. *Online Journal of Distance Learning Administration* *3*(4), 1-7.

Pierce, D., & Kinsell, S. (2008). *Cracking the TOEFL IBT 2008 edition*. The Princeton Review.

Pilliner, A. E. G. (1968). Subjective and objective testing. In Davies (Ed.). *Language testing symposium* (pp. 19-35). Oxford University Press.

Popham, W. J. (2001). *The truth about testing. An educator's call to action.* Association for supervision and curriculum development.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14. https://doi.org/10.3102/0013189X031007003

Powers, D. E. (2010*). The case for a comprehensive, four skills assessment of English language proficiency*. *TOEIC Compendium Study*, ETS Publications. https://www.ets.org/Media/Research/pdf/TC-10-12.pdf

Prapaisit de Segovia, L., & Hardison, D. M. (2008). Implementing education reform: EFL teachers' perspectives. *ELT Journal*, *63*(2), 154–162. https://doi.org/10.1093/elt/ccn024

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal*, *100*(1), 190–208. https://doi.org/10.1111/modl.12308

Randall, D. (2015). English studies in Turkey: An assessment. *Ariel: A Review of International English Literature*, *46*(1–2), 45–68. https://doi.org/10.1353/ari.2015.0013

Richards, J. C. (2002). *Planning aims and objectives in language programs*. Oxford Graphic Printers Pte Ltd.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialist in the assessment of criterion-referenced test-item validity. *Dutch Journal of Educational Research*, *2*, 49-60.

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement*, *33*(3), 271–290. https://doi.org/10.1111/j.1745-3984.1996.tb00493.x

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, *27*(2), 109–131. https://doi.org/10.1111/j.1745-3984.1990.tb00737.x

Scott, M., Stelzer, T., & Gladding, G. (2006). Evaluating multiple-choice exams in large introductory physics courses. *Physical Review Special Topics - Physics Education Research*, *2*(2), 020102: 1-14. https://doi.org/10.1103/PhysRevSTPER.2.020102

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple-choice question examination versus assignment essay. *Higher Education*, *35*, 453-472.

Scriven, M. (1967). The methodology of evaluation. In R. F. Stake (Ed.), *Curriculum evaluation: American Educational Research Association monograph series on evaluation, No. 1* (pp. 39-83). Rand McNally.

Seong, Y. (2017). Assessing L2 academic speaking ability: The need for a scenario-based assessment approach. *Studies in Applied Linguistics & TESOL, 17*(2), 36-40. https://doi.org/10.7916/salt.v17i2.1225

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Skehan, P. & Foster, P. (2001). The influence of task structure and processing conditions on narrative retellings. *Language Learning 49*(1), 93-120. https://doi.org/10.1111/1467-9922.00071

Stiggins, R. (2014). Improve assessment literacy outside of schools too. *Phi Delta Kappan*, *96*(2), 67–72. https://doi.org/10.1177/0031721714553413

Sugianto, A. (2016). An analysis of English national final examination for junior high school in terms of validity and reliability. *Journal on English as a Foreign Language, 6*(1), 31-42.

Sung, P.-J., Lin, S.-W., & Hung, P.-H. (2015). Factors affecting item difficulty in English listening comprehension tests. *Universal Journal of Educational Research*, *3*(7), 451–459. https://doi.org/10.13189/ujer.2015.030704

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, *10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Towns, M. H. (2010). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education*, *87*(1), 91–96. https://doi.org/10.1021/ed8000039

Valette, R. M. (1977). *Modern language testing.* Harcourt Jovanovich Publishers.

Wahyuni, A., Sukartingsih, L., & Herawati, A. (2014, October 7-9). *Blended learning in teaching reading: A pedagogical practice to teaching English as a foreign language in an Indonesian university context* (Paper presentation). The 61st TEFLIN International Conference, Solo, Indonesia.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*(2), 191-2018.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7–25. https://doi.org/10.1080/08957340709336728

Weber, R. P. (1990). *Basic content analysis*. Sage Publications.

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall.

Winke, P. M., & Isbell, D. R. (2017). Computer-assisted language assessment. In Thorne S., & May S. (Eds.), *Language, education and technology*. *Encyclopedia of language and education* (3rd ed., pp. 313-325). Springer.

Wu, M., Tam, H.P., & Jen, T.H. (2016). *Educational measurement for applied researchers*. Springer Nature.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5–21. https://doi.org/10.1177/0013164403258393