Developing a validity argument for the English Placement Listening Fall 2010 test at Iowa State University

by

Huong Le

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English to Speakers of Other Languages/Applied Linguistics

(Language Testing and Assessment)

Program of Study Committee

Volker Hegelheimer, Major Professor

John Levis

Denise Schmidt

Iowa State University

Ames, Iowa

2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The study was aimed at examining the usefulness of the English Placement Listening test (EPT) in Fall 2010 at Iowa State University (ISU) by using the current argument-based validation approach with a focus on four main inferences constructing the validity argument. Both qualitative and quantitative methods were employed. The results contributed both positive and negative attributes to the validity argument for the EPT Listening Fall 2010 test. The qualitative examination on the test specification and the test booklet showed that the test was authentic with a good distribution of question types and test item indices. In specific, the 30 test items were equally divided into comprehension and inference questions with 90% and 70% of them falling within an acceptable difficulty range, and an acceptable discrimination range respectively. General statistical analyses of the EPT Listening Fall 2010 test score set of 556 test takers produced a normal distribution with a reliability of nearly 0.70. Moreover, the correlation analyses among different set scores of the EPT Fall 2010 test takers supported the usefulness of the EPT test in discriminating proficiencies of the test takers besides their TOEFL scores. However, numerous weaknesses were detected such as an incomplete test specification, weak strengths of the correlational relationships between the EPT test and the TOEFL tests ($r<0.6$). The study provided an evidence on the importance of the operation of the EPT test at ISU and lead to some recommendations on supporting the validity argument for the test.

CHAPTER 1: INTRODUCTION

This chapter is to introduce the topic of my study, and present the main reasons for choosing it. After that, a close look at some questions that I would like to address within the scope of the study will be given. A brief overview of the following chapters will close the chapter.

Statement of problem

There are two main groups of forces that have driven me to look into the validity of the English Language Placement (EPT) Listening test at ISU. The first bases on my review of current validation theories or practice in language testing and assessment, which has helped me come up with some questions of interest to be researched. The second comes from my actual experiences with the EPT test at ISU that have intrigued me to carry out this study to examine the effectiveness and usefulness of the test.

Validity and validation in language testing and assessment

Considered to be the most important and complex concept in language testing, validity has been under examination by numerous testing experts and researchers, and has had its own life in the field of language testing and assessment (Chapelle, 1999; Kane, 2001). In the early 1960s, despite being described as an utmost characteristic of a language test (Lado, 1961, p. 321), validity was generally seen to connect with the test itself, and test scores (Bachman, 1990; Chapelle, 1999; Kane, 2001; Messick, 1989). A thorough examination into the definition of validity had not occurred until the early 1990s. The current view has revealed the complex nature of validity, which is a unified evaluation of the interpretation or use of test scores (APA, 1985; AERA et al., 1999; Bachman, 1990; Kane, 2001; Messick, 1989). Thus, the question of how the current view has shaped the testing and assessment practice has motivated me to do more theoretical and empirical research in order to have a proper and cynical insight into this concept.

Validation in language testing and assessment is generally explained as a process to investigate validity (AERA et al., 1999; Bachman, 1990; Chaplle, 1999; Messick, 1989); therefore, the evolution of the concept of validity in language testing and assessment has accompanied with changes in how to conceptualize the notion of validation. So far there have been two main approaches in validation studies including (1) accumulation-of-evidence

approach, and (2) argument-based approach (Chapelle, 1999; Kane, 2001). While the first approach sees the validation as a collection of evidences to support or refute a certain test score interpretation or use, the second approach views it as an on-going and critical process in order to build up a validity argument for a certain test.

One of the current models in the second approach, which has been much supported, employs the concept of interpretative argument in educational measurements proposed by Kane (1992, 2002, 2004). Accordingly, a validity argument of a certain test is built upon an interpretative argument constructed by logically ordered inferences, and a validity conclusion is viewed as an argument-based, context-specific judgment (Chapelle, 1999, p. 264). However, how to implement this approach in validation studies is another question, which requires more and more practical studies. A few of the latest validation studies in language testing and assessment have attempted to use this approach (Chapelle, Enright, & Jamieson, 2008; Chapelle, Jamieson, & Hegelheimer, 2003; Chapelle et al., 2010). The review of this interpretative argument-based validation literature and its relevant studies has given me another impetus to conduct a validation study using this latest approach.

Finally, the examination of relevant studies in language placement testing shows that a lot of efforts have been made in order to scrutinize different aspects of a placement testing, but the reliability and validity issues in language placement testing still call for more investigations and renovations despite its widespread use in institutions, universities or colleges. For example, some studies have looked at different instruments used for a language placement testing, or the ways to improve the quality of EPT (Brown, 1989; Sawyer, 1996; Wesche et al., 1993). Meanwhile, some researchers have been trying to address the issue of validity in placement testing (Brown, 1989; Fulcher, 1997; Goodbody, 1993; Lee, & Greene, 2007; Schmitz & DelMas, 1991; Truman, 1992; Usaha, 1997; Wall, Clapham, & Alderson, 1994). However, most of validation studies of EPT tests adopt the earlier accumulation-of-evidence validation approach in which different types of validity are examined separately for such a test (Fulcher, 1997; Schmitz & DelMas, 1991; Wall, Clapham, & Alderson, 1994).

These facts about language placement testing are good reasons for me to make an attempt to use the interpretative argument based model to examine an English placement testing at a university in the U.S.

The English Placement test (EPT) at Iowa State University (ISU)

With the annual high number of new international students, Iowa State University has employed EPT for a long time. The test is under the authority of the English Department, and is now supervised by Prof. Volker and Yoo-Ree. It is administered to all the international students admitted to the university whose native language is not English before each semester starts. It consists of three tests (Reading, Listening and Writing). In general, the goal of the test is to identify and assist the students who may face language problems to be successful in their academic studies; and the test results might influence their study plan, and budget for paying English courses. As a result, fair and accurate assessments of student abilities and decisions to assign individuals to appropriate English courses are very important to test-takers, and relevant test-users (English instructors, supervisors).

The two courses (519-Language testing and assessment, and language testing practicum-513) that I took in the last two semesters (Spring 2010, and Fall 2010), have given me valuable experiences with the EPT test at ISU, which have triggered me with some questions and strong motivations to investigate them.

First, despite its importance and quite long period in use, none of research has been carried out to evaluate the EPT test at ISU. This study is thus expected to be meaningful and practical to the test-users of the EPT test at ISU by giving some evidences on its usefulness. For instance, the study results will give some backing for or against their future decisions whether to maintain the test or not, and how to innovate it.

Secondly, I have had experiences with the EPT test at ISU in a number of roles as a test-taker, as an observer, or proctor, and as a test examiner for the test set used in the EPT Fall 2010 administration. Each of these various experiences has provided me with different biased evaluations or judgments about the plausibility of the test score interpretation and use. Thus, an empirical study will help me to address these hypotheses about the test.

Next, due to the limited scope of the study as a thesis project, I would like to narrow down the focus of the research onto the specific listening component of the EPT test at ISU in Fall 2010. In fact, my observation on the renovation of using authentic lectures with the integration of videos in the EPT Listening test has intrigued me to investigate the usefulness of the test.

The last not the least, with my deeply-rooted desire to develop an useful and good English placement test at my home university, this project is expected to bring me a profound insight into this specific area of interest, specifically using an argument-based validation approach in language placement testing, for my future professional development.

Statement of research questions

The research is aimed at structuring a validity argument for the use of the EPT Listening test at ISU, and then collecting some evidences supporting the argument based on the specific examination on its Fall 2010 administration. Based on the interpretative argument model proposed by Kane (2001; 2006) and exemplified in the article by Chapelle, Enright, Jamieson (2010), the first four inferences in the argument will be under investigation leading to four research questions in this study as following:

1. How do the EPT Listening test design and development help to measure what we want to measure of test-takers?
2. How reliable is the EPT Listening test in measuring test-takers' proficiencies respectively?
3. How do students' scores on the other test of language development (TOEFL) correlate with their scores on the EPT Listening test?
4. What are challenges to the validity argument of the EPT Listening test at ISU to be refuted?

Organization of the study

The study consists of five chapters. The first chapter – Introduction is aimed at introducing my topic area and giving main motivations for me to implement this project. The purpose of the second chapter – Literature review is to provide a profound theoretical and empirical background with a critical discussion on the relevant concepts, models, or theories for the study. Chapter 3 – Methodology gives a description on how the study is conducted accompanied with a review on each selected methodology. The following chapter – Chapter 4 is the presentation of the main results of the study and a discussion about them. Chapter 5 – Conclusion has three main aims, which are to summarize the main findings of the study, to specify the limitations of the study, and to suggest some directions for future investigations.

CHAPTER 2: LITERATURE REVIEW

This chapter consists of four main parts. The first two parts are aimed at theoretically and empirically examining the issue of validity in language testing and assessment, and the argument-based validation approach as a widely-supported approach. A critical review of how to put the argument-based validation approach into practice is the focus of the second part that begins with a cynical comparison of this approach with other approaches followed by a close look at the three latest validation studies employing this approach. The third part describes language placement testing, specifically English Placement testing (EPT) as an important type in language testing and assessment, and presents some concerns about how to investigate the validity of this testing type. These theoretical and empirical foundations act as driving forces leading to the restatement of the problems that will be addressed in my study in the fourth part.

1. Validation of a test in language testing and assessment

1.1. The conception of validity in language testing and assessment

What is validity?

Three important milestones in the conception of the current validity in language testing and assessment could be given here.

First, Messick (1989, p. 13) states that "validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment". Different from the earlier view about validity, this statement means that neither the test itself nor test scores per se is validated, but the interpretation determined by the proposed use is validated. Moreover, validity cannot be proved, but only be judged by the availability of theoretical rationales or empirical evidences.

Messick's view about validity was then supported and found an official recognition so that in the Standards for Educational and Psychological Testing (1985), validity is described as follows:

> The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept (APA, 1985, p. 9).

This definition is well-explained and elaborated by Bachman (1990). First, in concert with the Messick's view, this definition helps to confirm that the inferences made on the basis of test scores, and their uses are the object of validation rather than the tests themselves. Second, according to him, validity has a complex nature comprising of a number of aspects including content validity, construct validity, concurrent validity, and consequences of test use; however, validity should be considered as a unitary concept pertaining to test interpretation and use with construct validity as an overarching validity concept. The synthesis of these explanations on the concept of validity in testing and assessment lead to the restatement of validity in the Standards for Educational and Psychological Testing in 1999. It states that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 1999, p. 9).

In a thorough examination of these statements about validity (AERA et al., 1999; APA, 1985; Bachman, 1990; Messick, 1989). Kane (2001) reveals four important aspects of this current view. First, validity involves an evaluation of the overall plausibility of a proposed interpretation or use of test scores. Second, consistent with the general principles growing out of construct validity, the current definition of validity (AERA et al., 1999; Messick, 1989) incorporates the notion that the proposed interpretations will involve an extended analysis of inferences and assumptions which includes both a rationale for the proposed interpretation and a consideration of possible competing interpretations. The resulting evaluative judgment reflects the adequacy and appropriateness of the interpretation and the degree to which the interpretation is adequately supported by appropriate evidence. Fourth, validity is an integrated, or unified, evaluation of the interpretation; and it is not simply a collection of techniques, or tools.

Different aspects of validity

In recognition of the complexity of validity and its importance in test evaluation, a number of aspects of validity have been examined (Bachman, 1990; 2004; Bachman & Palmer, 1996; Brown, 1996). Based on the concept of test use, both Bachman (1990, p. 243) and Brown (1996, p. 233) agree on the three main aspects of validity: content relevance and content coverage (or content validity), criterion relatedness (or criterion validity), and meaningfulness of construct (or construct validity). In addition, as discussing testing in language programs, Brown (1996, p. 249) suggests the examination of standards setting or the appropriateness of a cut-point as another important aspect of validity.

First, content validity involves characteristics of a test itself, not test score interpretations and use (Bachman, 1990, p. 243; Brown, 1996, p. 232). There are two aspects of content of a test under examination for validity including content relevance and content coverage. Content relevance requires 'the specification of the behavioral domain in question and the attendant specification of the task or test domain (Messick, 1989, p. 117) as well as the specification of both the ability domain, and test method facets (Bachman, 1990, p. 244). On the other hand, content coverage is to examine the extent to which the tasks required in the test adequately represent the behaviorial domain in question, for instance, how test tasks are sampled to be representative from the domain.

Second, criterion validity (Bachman, 1990, p. 248-253; Brown, 1996, p. 246) refers to evidence involving a relationship between test scores and some criterion which is believed to be an indicator of the ability tested. This 'criterion' may be level of ability as defined by group membership, individuals' performance on another test of the ability in question, or their relative success in performing tasks that involves this ability. There are two types of criterion relatedness: (1) concurrent validity and (2) predictive validity. Concurrent validity studies are purposed to examine differences in test performance among groups of individuals at different levels of language ability, and to examine correlations among various measures of a given ability. However, predictive validity is to provide information on how well test scores predict some future behavior by carrying a correlation study demonstrating the relationship between test-takers' scores on the test and their actual performance.

Construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of constructs (Bachman, 1990, p. 254; Brown, 1996, p. 239). Thus, construct validation seeks to provide both logical analysis and empirical evidence that support specific inferences about relationships between constructs and test scores. First, logical analysis is involved in defining the constructs theoretically and operationally while empirical evidence supporting construct validity comprises of several types (1) the examination of patterns of correlations among item scores and test scores, and between characteristics of items and tests and scores on items and tests, (2) analyses and modeling of the processes underlying test performance, (3) studies of group differences, (4) studies of changes over time, or (5) investigation of the effects of experimental treatment.

Finally, standard setting that is defined as the process of deciding where and how to make cut-points, provides an important evidence on validity of testing in a certain language program (Brown, 1996, p. 249). Its importance lies in the fact that setting standards of performance is basically used for making five important types of decisions in language programs: (1) admitted into an institution, (2) placed in the elementary, intermediate, or advanced level of a program, (3) diagnosed as knowing certain objectives or not knowing others, (4) passed to the next level of study, or (5) certified as having successfully achieved the objectives of a course or program.

1.2. Approaches in validation studies in language testing and assessment

1.2.1. The concept of validation in language testing and assessment

Some explanations about validation based on the latest view on validity (AERA et al., 1999; Mesick, 1989) will be presented here. First, Bachman (1990, p. 96) explains that validation is "a process through which a variety of evidence about test interpretation and use is produced; such evidence can include but is not limited to various forms of reliabilities and correlations with other tests." Likewise, in the Standards for Educational and Psychological Testing, the concept of validation is described as follows "validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA et al., 1999, p. 9).

Significantly, based on the latest view about validity, the validation process is expanded to be seen as an on-going procedure in the life cycle of a test with the integration of test impact as an aspect of validity. In specific, "validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. As validation proceeds and new evidence about the meaning of a test's scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test" (AERA et al., 1999, p. 9). Also, in cases where test-based decisions have serious consequences, validation involves evaluating the full, decision-based interpretations, and not just the descriptive interpretations on which the decision is based. Hence, Kane (2001) supports that validation involves the evaluation of the credibility of an interpretation per se, and its role in evaluating the legitimacy of a particular use.

What is noticeable about the examination of the validity process of a certain test is the identification of the roles of the two parties involved including test developer and test user. Accordingly, "the test developer is responsible for furnishing relevant evidence and a rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used" (AERA et al., 1999, p. 11). In other words, while those who develop a test are responsible for giving relevant and plausible theoretical and empirical backing for using the test for certain purposes, those who propose to use a test score in a particular way are expected to justify this use by showing that the positive consequences of the proposed use outweigh the anticipated negative consequences.

1.2.2. Main approaches in validation studies in language testing and assessment

In recognition of the interrelationship between the concepts of validity and validation, it is logical to find that how to conduct a validation study is influenced by how these two concepts are viewed (Bachman, 1990; Chapelle, 1999; Kane, 2001).

Chapelle (1999) and Kane (2001) attempted to give a brief summary of main approaches in validation studies based on the history of the validity concept in language testing and assessment. According to them, two main approaches in validation research in language testing can be noticed (1) accumulation-of-evidence approach, and (2) argument-based approach.

The first approach or the accumulation-of-evidence approach derives from the past work in educational measurement (Cronbach & Meehl, 1955; Messick, 1989) and language assessment (Bachman, 1990; Weir, 2005). This approach sees the final result of validation or a validity conclusion more a proof-based, and categorical result (Chapelle, 1999, p. 264). In specific, the investigation into validity is simply to collect and present different evidences on different aspects of validity such as reliability, and construct validity (Chapelle, 1999, p. 258). As described by Kane (2001), this period witnessed the development of three major validation models in correspondence with the three main aspects of validity including the criterion-based model, the content-based model, and the construct-based model. A number of validations studies in language testing and assessment have employed this approach (Brown, 1989; Fulcher, 1997; Lee, & Greene, 2007; Schmitz & DelMas, 1991; Truman, 1992; Usaha, 1997; Wall, Clapham, & Alderson, 1994).

On the other hand, by the late 1980s, the second approach was born in order to solve the issue of selecting and synthesizing different sources in making a proper judgment on validity by

using a consistent framework for structuring these sources in terms of arguments (Cronbach, 1990, 1988; Toulmin et al., 1979). In specific, they call for a view on validity as an evaluative argument with relevant social dimensions and contexts of using a test and a structure for the analysis and presentation of validity data. This view has been developed and received supports from many researchers since then (Cronbach, 1988; Crooks, Kane, & Cohen, 1996; Kane, 1992; Shepard, 1993). The latest argument-based validation model receiving a lot of supports is proposed by Kane (1992; 2001; 2002). The model is based on Messick's (1989) conception of validity with his outline of validity evidence types, the concept of interpretative argument in educational measurements proposed by Kane (1992, 2002, 2004). Different from the first approach, it views a validity conclusion as an argument-based, context-specific judgment (Chapelle, 1999, p. 264). This approach have been illustrated in several recent validation studies in language testing and assessment (Chapelle, Enright, & Jamieson, 2008; Chapelle, Jamieson, & Hegelheimer, 2003; Chapelle et al., 2010).

On the whole, a comparative view between the accumulation-of-evidence approach and the argument-based approach can be summarized here. The validation process entails providing a number of relevant theoretical rationales and empirical evidence. In other words, it calls for the researcher and any test-user to draw on multiple sources of information to create an integrated, multifaceted evaluation where a language test is concerned, rather than basing it on a single research result or set of results. However, according to a number of researchers in testing and assessment (Bachman, 1990; Chapelle, 1999; Kane, 1992, 2001, 2002, 2004), the accumulation-of-evidence approach can be problematic because of the difficulty in deciding what kind of evidence to gather and how much evidence is enough. On the other hand, the argument-based approach has more advantages. For example, it emphasizes that validity is not a yes or no answer, but is contextually-based without an ending point.

2. The argument-based validation approach in language testing and assessment

2.1. Using interpretative argument in examining validity in language testing and assessment

The argument-based validation approach in language testing and assessment views validity as an argument construed by an analysis of theoretical and empirical evidences instead of a collection of separately quantitative or qualitative evidences (Bachman, 1990; Chapelle, 1999; Chapelle et al., 2008, 2010; Kane, 1992, 2001, 2002; Mislevy, 2003). One of the widely-

supported argument-based validation frameworks is to use the concept of interpretative argument (Kane, 1992; 2001; 2002). This approach is clearly defined in his article 'An argument-based approach to validity' as follows:

> The argument-based approach to validation adopts the interpretative argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. (Kane, 1992, p. 527)

Some explanations for using interpretative arguments to examine validity in language testing and assessment can be made here. First, validity is associated with the interpretation assigned to test scores (AERA et al., 1999; Bachman, 1990; Chapelle, 1999; Messick, 1989). Moreover, the interpretation assigned to test scores involves an argument leading from the scores to score-based statements or decisions. This means that the assumptions inherent in the proposed interpretations and uses of test scores can be made explicit in the form of an interpretative argument that lays out the details of the reasoning leading from the test performances to conclusions included in the interpretation and to any decisions based on the interpretation.

Therefore, in the light of the argument-based approach, validity cannot be proved, but depends on the plausibility of interpretative arguments that can be critically evaluated with evidence. Moreover, the kinds of evidence needed for the validation of a test-score interpretation can be identified systematically by an explicit recognition of the inferences or assumptions or the details in the interpretative arguments.

2.2. Conducting an argument-based validation study in language testing and assessment

A number of attempts have been made on how to build a validity argument in language testing and assessment using the concept of interpretative argument (Bachman, 1990; Chapelle, 1999; Kane, 1992, 2001, 2002). First, Kane (1992, p. 534) asserts that **"**the argument-based approach to validity is basically quite simple. One chooses the interpretation, specifies the interpretative argument associated with the interpretation, identifies competing interpretations, and develops evidence to support the intended interpretation and to refute the competing interpretations. The amount of evidence and the types of evidence needed in a particular case depend on the inferences and assumptions in the interpretative argument". Likewise, based on Messick's (1989) guidelines and Shepard's (1997) explanations, Chapelle explains how to conduct argument-based validation studies (1999, p. 258-265). Validation begins with a

hypothesis about the appropriateness of testing outcome, which refers to assumptions about what a test measures and what their scores can be used for. Such hypotheses may be developed from testing or construct theories, or anticipated testing consequences such as test-takers' emotions after the test. Next will be the collection of relevant evidence for testing the hypotheses. Data pertaining to the hypothesis are gathered, and results are organized into an argument from which a "validity conclusion" can be drawn about the validity of testing outcomes.

Based on Kane's concept of interpretative argument and Mislevy's description about assessment as reasoning from evidence, Bachman gives a framework for the argument-based validation process consisting of two main steps: articulating a validation argument, and collecting different kinds of evidence in support of a validation argument (Chapter 9, 1990). The first step has two main functions: (1) to provide a guide for the process of designing and developing tests, and (2) to provide a framework for collecting evidence in support of the intended interpretations and uses. For the second step, he suggests some different types of evidences in order to support the validity argument. They include quantitative evidences such as carrying out descriptive statistical analyses, or correlation analyses, and qualitative evidences like the analysis of test content, the analysis of test-taking processes, the analysis of correlations among scores from a large number of tests, the analysis of differences among non-equivalent criterion groups.

In the following paper, Kane (2001, p. 330) outlines some strategies for validating the test score interpretation, and expands the validation process as an on-going cycle. The main steps in the validation cycle of a test can be presented as below:

> (1) State the proposed interpretative argument as clearly and explicitly as possible.
>
> (2) Develop a preliminary version of the validity argument by assembling all available evidence relevant to the inferences and assumptions in the interpretative argument. One result of laying out the proposed interpretations in some detail should be the identification of those assumptions that are most problematic.
>
> (3) Evaluate empirically and/or logically the most problematic assumptions in the interpretative argument. As a result of these evaluations, the interpretative argument may be rejected, or it may be improved by adjusting the interpretation and/or the measurement procedure in order to correct any problems identified.

(4) Restate the interpretative argument and the validity argument and repeat Step 3 until all inferences in the interpretative are plausible, or the interpretative argument is rejected.

2.3. Building a validity argument in language testing and assessment

Interpretative argument vs. validity argument

In the discussion about how to utilize the argument-based approach in validation studies in language testing and assessment, Kane (2001, p. 180) recommends the drawing of a distinction between an interpretative argument and a validity argument. Accordingly, the interpretative argument is to provide an explicit statement of the reasoning leading from test performances to conclusions and decisions; on the other hand, the validity argument provides an evaluation of the plausibility of the interpretative argument.

Interpretative argument

What is an interpretative argument?

In the article about how to put the argument-based approach in practice, Kane (2002) summarizes the common description of an interpretative argument agreed by various testing researchers (Crooks, Kane, & Cohen, 1996; Kane, 1992; Shepard, 1993). It states that "an interpretative argument is known as a network of inferences and supporting assumptions leading from scores to conclusions and decisions" (Kane, 2002, p. 231)

Based on rationales about kinds of arguments and structures of arguments (Toulmin et al., 1979), Kane (1992; 2002) attempts to explain an interpretative argument as a type of practical arguments which address issues in various disciplines and in practical affairs. In practical arguments, "because the assumptions cannot be taken as given and because the available evidence is often incomplete and, perhaps, questionable, the argument, is, at best, convincing or plausible. The conclusions are not proven" (Kane, 1992, p. 527). Therefore, Kane (1992, 2002) points out that, unlike purely logical or mathematical arguments, the assumptions in an interpretative argument cannot be taken as given, and the evidence in support of these assumptions is often incomplete or debatable. Thus, the conclusions of interpretative arguments are not proven, but can only be evaluated in terms of how convincing or plausible they are. He also presents three criteria for evaluating the inferences made on the basis of an interpretative argument: (a) clarity of argumentation, (b) coherence of argument, and (c) plausibility of assumptions. The first characteristic means that the argument should be stated clearly so that

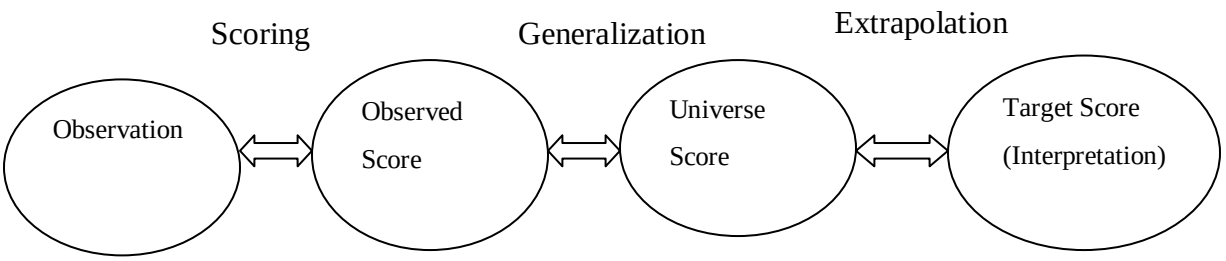
what it claims and what it assumes are known. Next, the coherence of an interpretative argument refers to the logic and reasonability of the conclusions given the assumptions. Third, the assumptions should be plausible or supported by evidence. Sources of evidence can include parallel lines of evidence or plausible counterarguments to refute.

Structure of an interpretative argument

A lot of testing researchers have been interested in examining different kinds of inferences construing an interpretative argument (Bachman, 2004; Crooks, Kane & Cohen, 1996; Kane, Crooks & Cohen, 1999; Kane, 2002). Crooks, Kane and Cohen (1996) have identified several commonly found inferences in test-score interpretations. Five of these inferences are evaluation, generalization, extrapolation, explanation, and decision-making, each of which requires a different mix of supporting evidence. In a close examination of the nature of interpretative argument to validate high-stakes testing programs, Kane (2002, p. 33) categorize these inferences and assumptions into two broad categories: semantic and policy. The semantic inferences are those that lead from scores to conclusions or from one conclusion to another and are represented by the first four of the five kinds of inferences: evaluation, generalization, extrapolation, and explanation. They make claims about what the test scores mean. Policy inferences lead from conclusions to decisions and therefore involve the adoption of decision rules. The justification of such policies is generally based on claims that the decision rule will achieve certain desirable outcomes, and cause little or no negative impacts.

Kane, Crooks and Cohen (1999) attempted to illustrate how to structure an interpretative argument for a validation study of performance assessment. Accordingly, the development of the interpretative argument for performance assessment involves three inferences: scoring, generalization and extrapolation. The structure of the interpretative argument is illustrated in Figure 1.

Figure 1: Links in an interpretative argument (Kane, Crooks, & Cohen, 1999, p. 9)



In the figure, the argument consists of four parts each of which is linked to the next one by an inference. The first link - 'scoring' is an inference from an observation of performance to a score, and is based on the assumptions about the appropriateness and consistency of the scoring procedures and the conditions under which the performance is obtained. The second link – 'generalization' is from an observed score on a particular measure to a universe score, or the score that might be obtained from performances on multiple tasks similar to those included in the assessment. This link is based on the assumptions of measurement theory. The third link – 'extrapolation' is from the universe score to a target score, which is essentially an interpretation of what a test taker knows or can do, based on the universe score. This link relies on the claims in an interpretative argument and the evidence supporting these claims.

Validity argument

What is a validity argument?

Based on the concept of a validity argument and an interpretative argument given by a number of testing researchers (Cronbach, 1988; Kane, 1992, 2002; Messick, 1989), a validity argument is claimed to provide an overall evaluation of the plausibility of the proposed interpretations and uses of test scores. It aims for a cogent presentation of all of the evidence relevant to proposed interpretations, and to the extent possible, the evidence relevant to plausible alternate interpretations. Therefore, how to structure a validity argument has intrigued researchers in language testing and assessment in order to address the concerns about judging its plausibility and ensuring the consistency in using the argument-based approach in validation studies in this field.

Structure of a validity argument in a validation study in language testing and assessment

The construction of a validity argument is suggested to base on Toulmin's (2003) argument structure. According to Toulmin, an argument consists essentially of claims made on the basis of data and warrants. The structure of the argument is illustrated by Bachman (2004, p. 9) and can be found below (see Figure 2).

Some explanations for each component of the structure of a validity argument can be given here. In this description, a claim is "a conclusion whose merits we are seeking to establish" (Toulmin, 2003, p. 90). In other words, a claim is the interpretation that we want to make on the basis of the data, about what a test taker knows or can do. Next, data includes "information on which the claim is based" (Toulmin, 2003, p. 90). For example, in the case of testing and

assessment, these are the responses of test-takers to assessment tasks, or what test takers say or do as taking the test. Finally, warrants and rebuttals act as a link between data and a claim, and are carefully examined in terms of their nature and structure (Toulmin, 2003, p. 91). A warrant is defined as a general statement that provides legitimacy of a particular step in the argument (Toulmin, 2003, p. 92). As being seen in Figure 2, the arrow from the data to the claim represents an inference, which is justified on the basis of a warrant. Warrants are thus propositions that we use to justify the inference from data to claim. For example, it can be a deduction that students who are able to support character descriptions with specifics will do so in tasks like the one at hand. Moreover, the justification of a warrant is based on backing. The backing is explained to include "other assurances, without which the warrants themselves would posses neither authority nor currency" (Toulmin, 2003, p. 96). On the other hand, a rebuttal consists of "exceptional conditions which might be capable of defeating or rebutting the warranted conclusion" (Toulmin, 2003, p. 94). As can be understood from the definition, rebuttals present counterclaims or alternative explanations to the intended inference, and the rebuttal data consist of evidence that may support, weaken, or reject the counterclaims.

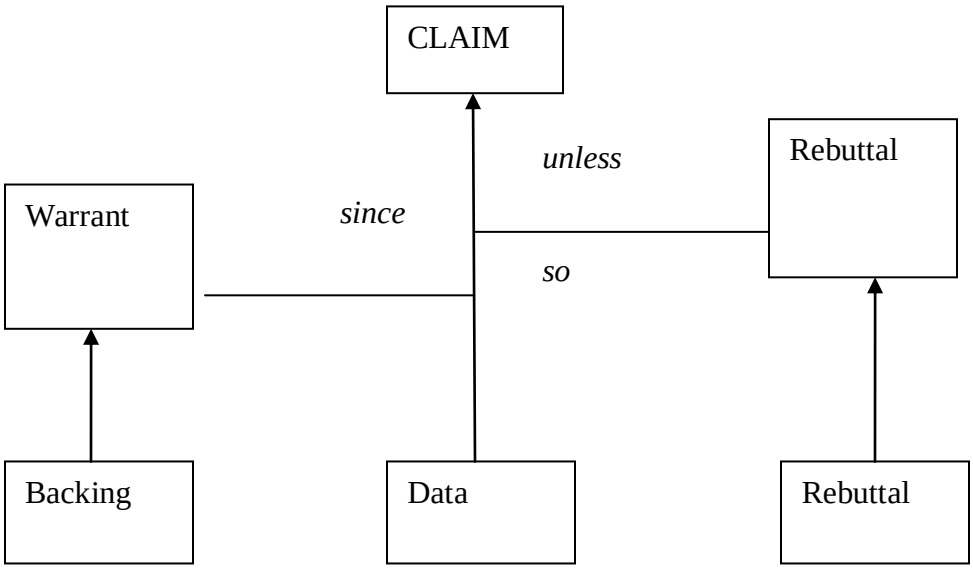


Figure 2: Toulmin's diagram of the structure of arguments (taken from Bachman, 2004, p. 9)

Sources of backing for warrants in a validity argument

Many researchers have made great contributions to how to collect evidences to back a warrant in a validity argument for a certain test. Six main sources which are frequently used in validation studies in language testing and assessment are content analysis, empirical item or task

analysis, dimensionality analysis, relationships of test scores with other tests and behaviors, research on differences in test performance, and arguments on testing consequences or washback studies (Bachman, 2004; Chapelle, 1999; Kane, 2002; Mislevy et al., 2003).

Due to the existence of various sources of backing for warrants in a validity argument, two main strategies on how to select which source of backing are recommended. In specific, Kane (2002, p. 32) emphasizes the importance of the proposed interpretation in deciding which kind of evidence is required for validation, and the variability in plausibility of a validity argument for one test in different contexts, or different populations of examinees. According to him, it is entirely possible for one or more of these interpretations to be valid, where other interpretations are invalid. For example, it is possible that the test scores provide a good indication of an examinee's skill in solving the kind of problem included in the test, but provide a poor indication of skills in any wider set of problems or in any other context.

2.4. A critical review of the argument-based validation approach

The argument-based approach employing an interpretative argument offers several advantages, and presents some current concerns to solve.

A major strength of this argument-based approach to validation is the guidance it provides in allocating research efforts and in deciding on the kinds of validity evidence that are needed (Bachman, 2004; Cronbach, 1988; Kane, 1992, p. 335-337). First, the structure of the interpretative argument determines the kinds of evidence to collect at each stage of the validation effort and provides a basis for evaluating overall progress. Kane (1992, p. 535) explains that this approach does not identify any kind of validity evidence as being generally preferable to any other kind of validity evidence, but the selection of validity evidence should address the plausibility of the specific interpretative argument being proposed. For instance**, t**he kinds of validity that are most relevant are those that evaluate the main inferences and assumptions in the interpretative argument, particularly those are most problematic. And the weakest parts of the interpretative argument are to be the focus of the analysis. Moreover, if some inferences in the argument are found to be inappropriate, the interpretative argument needs to be either revised or abandoned.

Second, Kane (1992, p. 535) emphasizes that the evaluation of an interpretative argument does not lead to any absolute decision about validity, but it does provide a way to gauge progress. In other words, it views the validation of a certain testing as an on-going and critical

process instead of a static process with a clear answer of either ‘valid’ or ‘invalid’. As the most questionable inferences and assumptions are checked, and either are supported by the evidence or are adjusted so that they are more plausible, the plausibility of the interpretative argument as a whole can improve. For instance, if evidences from this evaluation of the validity argument indicate that there exists a problem in some specific aspects of measurement procedures, some ways to solve the problem and thereby to improve the procedure will be suggested. Moreover, the criticism and thoroughness of this approach can be seen through its recognition of the role of an audience as the subjective to be persuaded, the need to develop a positive case for the proposed interpretation, and the need to consider and evaluate competing interpretations. For example, through exploring the validation of such tests, readers can gain an insight into the main steps of developing the tests, and judge the validity argument of the tests based on theoretical backgrounds, as well as empirical evidence provided.

Significantly, these two main advantages of using interpretative arguments in the argument-based validation approach in language testing and assessment are well-illustrated in an insightful discussion based on real experiences of the testing researchers as implementing the project of building a validity argument for the test of English as a foreign language (TOEFL) developed by the English Testing Service (ETS) (Chapelle, Enright, & Jamieson, 2010). The discussion clearly points out the difference in approaching validity of a test by employing the interpretative argument-based approach suggested by Kane (1992, 2002).

However, there are still some concerns with how to put the argument-based approach into practice in language testing and assessment. First, Bachman (2004) claims that the interpretative argument-based validation approach (Kane, 1992, 2002; Kane, Crooks, & Cohen, 1999; Mislevy et al., 2003) has not yet addressed the issue of test impact as an aspect of test validity in language testing and assessment. He points out that a framework based on the argument-based validation approach provides a logic set of procedures for investigating and supporting claims about score-based inferences, but still fails to include the claims about test use and its consequences. This issue should be addressed in validation studies as using the interpretative argument. Second, after reviewing relevant validation studies in language testing and assessment, another issue with the argument-based approach is the lack of a systematic framework and guidelines in order to assure the consistency among validation studies employing this approach. Also, few validation studies have examined mid-stakes or low-stakes tests which are in fact very popular in language

programs (Brown, 1996). Therefore, more efforts should be made in order to guide how to use the argument-based approach to examine validity of such tests.

2.5. The argument-based validation approach in practice so far

Several recent validation studies in language testing and assessment have attempted to take the argument-based approach into practice, three of which are chosen to be illustrated here (Chapelle, Enright, & Jamieson (2008); Chapelle, Jamieson, & Hegelheimer (2003); Chapelle et al., 2010). The first one carried out by Chapelle, Jamieson and Hegelheimer exemplifies the employment of the concept of test purpose (Shepard, 1993) to identify sources of validity evidence and the framework of test usefulness (Bachman & Palmer, 1996) to structure their validity argument. On the other hand, the other two illustrate the application of the structure of an interpretative argument to guide the validation process and to build a validity argument for the tests under examination.

The presentation of these three studies has some purposes. First, it is aimed at visualizing how to put the argument-based validation approach into practice which acts as an empirical foundation for my study. Second, it is expected to help understand the advantages of using the concept of interpretative argument to address some aforementioned concerns in the argument-based validation approach including: (1) involving impacts addressed through decisions made during the course of design and the initial validation of an ESL test, (2) providing guidelines on how to use the argument-based approach in examining validity in language testing and assessment such as identifying relevant theories or types of evidences, and (3) developing and judging the plausibility of a validity argument for different kinds of tests (high-stakes, mid-stakes, or low-stakes).

(1) Validation of a web-based ESL test (Chapelle, Jamieson, & Hegelheimer, 2003)

In the study by Chapelle, Jamieson, and Hegelheimer (2003), the researchers exemplified the use of the argument-based approach through the validation of a web-based ESL test – a low-stakes type. The validity argument for the test was critically built by employing the concept of test purpose (Shepard, 1993), and the notion of test usefulness (Bachman & Palmer, 1996). The test under investigation is a part of a web-based language system that is aimed at offering an interactive language learning activities for English language learners. The test called *Test Your English (TYE)* was developed over an eight-month period in 2000-01. The test results will be used to direct learners to the appropriate parts of the website for practicing their English.

A number of steps in building up a validity argument for the web-based ESL test were taken in the study. First, the researchers carefully described the original purpose, design and development of the test in order to explain how the test purpose influenced some main test-related decisions. Then, the validity argument was developed as comprising of both positive and negative theoretical and empirical attributes structured under six main characteristics in the framework of test usefulness given by Bachman and Palmer (1996). The six characteristics are (1) reliability, (2) construct validity, (3) authenticity, (4) interactiveness, (6) impact, (7) practicality.

The study is a good attempt to illustrate how to apply the current argument-based validation theory to develop a low-stakes, web-based ESL assessment. In specific, the study helps to answer three main questions regarding complexities in developing a validity argument. First, it helps to give an answer to what kinds of theoretical rationales can be brought to bear on a validity argument. The study demonstrates how a number of theoretical rationales can be used to develop a means for articulating data analysis procedures that would test the data fit to construct theory, or construct validation. For example, theories of text difficulty and item difficulty underlay the design of the different level tests and the strategy of comparing item difficulty across level tests, or theories of vocabulary and grammatical development form the basis for item selection and analysis. Second, the study shreds some light on the question about how to take testing consequences into account as one aspect of validity. Specifically, in the study, the authors explain the integration of the intended impact as part of the test purpose into the design and development of the test as an evidence supporting its validity argument. Next, with the proposal of using the framework of usefulness to structure a validity argument, the study suggests a way to organize relevant sources of evidences in order to evaluate the validity argument. To be specific, the authors organize both positive and negative attributes under each characteristic of test usefulness, which can be either theoretical or empirical evidences as well as counterarguments to refute. The construction of the validity argument in the study also emphasizes the view of validation as a continual and cynical process. Accordingly, the negative attributes help to pave a way for additional steps to improve the test.

(2) Building a validity argument for the TOEFL (Chapelle, Enright, & Jamieson, 2008)

Different from the earlier validation study of a web-based test by Chapelle, Jamieson, and Hegelheimer (2003), the researchers employ and systematically develop Kane's

conceptualization about an interpretative argument in order to build a validity argument for the TOEFL test (Chapelle, Enright, & Jamieson, 2008). The whole project comprises of detailed descriptions about the interpretative argument for the TOEFL, a collection of relevant theoretical and empirical evidences on different aspects of validity of the test, and a construction of the validity argument for the TOEFL. The main components of the interpretative argument and the validity argument are illustrated in Table 1 and Figure 3 respectively.

Table 1: Summary of the inferences, warrants in the TOEFL validity argument with their underlying assumptions (Chapelle, Enright, Jamieson, 2010, p. 7)

Inference	**Warrant Licensing the Inference**	**Assumptions Underlying Inferences**
Domain description	Observations of performance on the TOEFL reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education.	1. Critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified. 2. Assessment tasks that require important skills and are representative of the academic domain can be simulated.
Evaluation	Observations of performance on TOEFL tasks are evaluated to provide observed scores reflective of targeted language abilities.	1. Rubrics for scoring responses are appropriate for providing evidence of targeted language abilities. 2. Task administration conditions are appropriate for providing evidence of targeted language abilities. 3. The statistical characteristics of items, measures, and test forms are appropriate for norm-referenced decisions.
Generalization	Observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters.	1. A sufficient number of tasks are included in the test to provide stable estimates of test takers' performances. 2. Configuration of tasks on measures is appropriate for intended interpretation. 3. Appropriate scaling and equating procedures for test scores are used. 4. Task and test specifications are well defined so that parallel tasks and test forms are created.
Explanation	Expected scores are attributed to a construct of academic language proficiency.	1. The linguistic knowledge, processes, and strategies required to successfully complete tasks vary across tasks in keeping with theoretical expectations. 2. Task difficulty is systematically influenced by task characteristics. 3. Performance on new test measures relates to performance on other test-based measures of language proficiency as expected theoretically. 4. The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components. 5. Test performance varies according to the amount and quality of experience in learning English.
Extrapolation	The construct of academic language proficiency as assessed by TOEFL accounts for the quality of linguistic performance in English-medium institutions of higher education.	Performance on the test is related to other criteria of language proficiency in the academic context.
Utilization	Estimates of the quality of performance in the English-medium institutions of higher education obtained from the TOEFL are useful for making decisions about admissions and appropriate curricula for test takers.	1. The meaning of test scores is clearly interpretable by admissions officers, test takers, and teachers. 2. The test will have a positive influence on how English is taught.

As can be seen, the interpretative argument for the TOEFL consists of six main different inferences (domain description, evaluation, generalization, explanation, extrapolation, and utilization), each of which consists of corresponding warrants and assumptions. These six inferences then prompt particular investigations throughout the process of research and development of the TOEFL iBT in order to construct the validity argument.

First, the domain description is built on the warrant that observations of performance on the TOEFL reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education. This warrant, in turn, is based on the assumptions (a) that assessment tasks representing the academic domain can be identified, (b) that critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified, (c) that assessment tasks requiring important skills and representing the academic domain can be simulated as test tasks. Some instruments used to support these assumptions are domain analysis, simulation of academic tasks, which help to bridge the inference from the target-language-use domain to relevant, and observable performances. Some examples used to support the domain description inference in the interpretative argument for the TOEFL are reports that (a) examine the nature of professional knowledge about academic language proficiency, (b) survey language tasks in an academic context, (c) report empirical investigations of students' and teachers' views about academic language tasks.

Evaluation means that observations of performance on TOEFL tasks are evaluated to provide observed scores reflective of targeted language abilities. This warrant is based on three assumptions about scoring and conditions of task administration: (a) rubrics for scoring responses are appropriate for providing evidence of targeted language abilities, (b) task administration conditions are appropriate for providing evidence of targeted language abilities, and (c) the statistical characteristics of items, measures, and test forms are appropriate for norm-referenced decisions. Accordingly, the relevant studies backing these assumptions will focus on appropriate scoring rubrics, task administration conditions, and psychometric quality of norm-referenced scores. For example, in order to support the assumption on task administration condition for the TOEFL Listening with the permission of note-taking, the study result showing that listening ability was determined to be elicited best through the use of tasks that provided test takers with opportunities to take notes, was given. Likewise, the psychometric results from the

TOEFL iBT field study were reported to provide good backing for the psychometric quality of the scores.

Generalization is based on the warrant that observed scores are considered as estimates of expected scores that test takers would receive on comparable tasks, test forms, administrations, and rating conditions. Four assumptions are identified to underly this warrant: (a) a sufficient number of tasks are included on the test to provide stable estimates of test takers' performances, (b) the configuration of tasks on measures is appropriate for the intended interpretation, (c) appropriate scaling and equating procedures for test scores are used, and (d) task and test specifications are well-defined so that parallel tasks and test forms are created. Consequently, some sources for backing these assumptions can be obtained from reliability analyses.

The explanation inference is built on the warrant that expected scores are attributed to a construct of academic language proficiency. Five assumptions about the construct of language proficiency are identified to underly this warrant, and are explained to rely on perspectives towards construct definition, or explanations for performance consistency: (1) test performance varies according to the amount and quality of experience in learning English, (2) performance on new test measures relates to performance on other test-based measures of language proficiency as expected theoretically, (3) the internal structure of the test score is consistent with a theoretical view of language proficiency as a number of highly interrelated components, (4) the linguistic knowledge, processes, and strategies required to successfully complete tasks vary in accordance with theoretical expectations, (5) task difficulty is systematically influenced by task characteristics. Thus, these assumptions can be supported by the results from (1) an examination of task completion processes and discourse for specific tasks, (2) correlation studies among TOEFL measures and other tests, (3) correlation analyses among measures within the TOEFL test, (4) research about expected relationships with English learning.

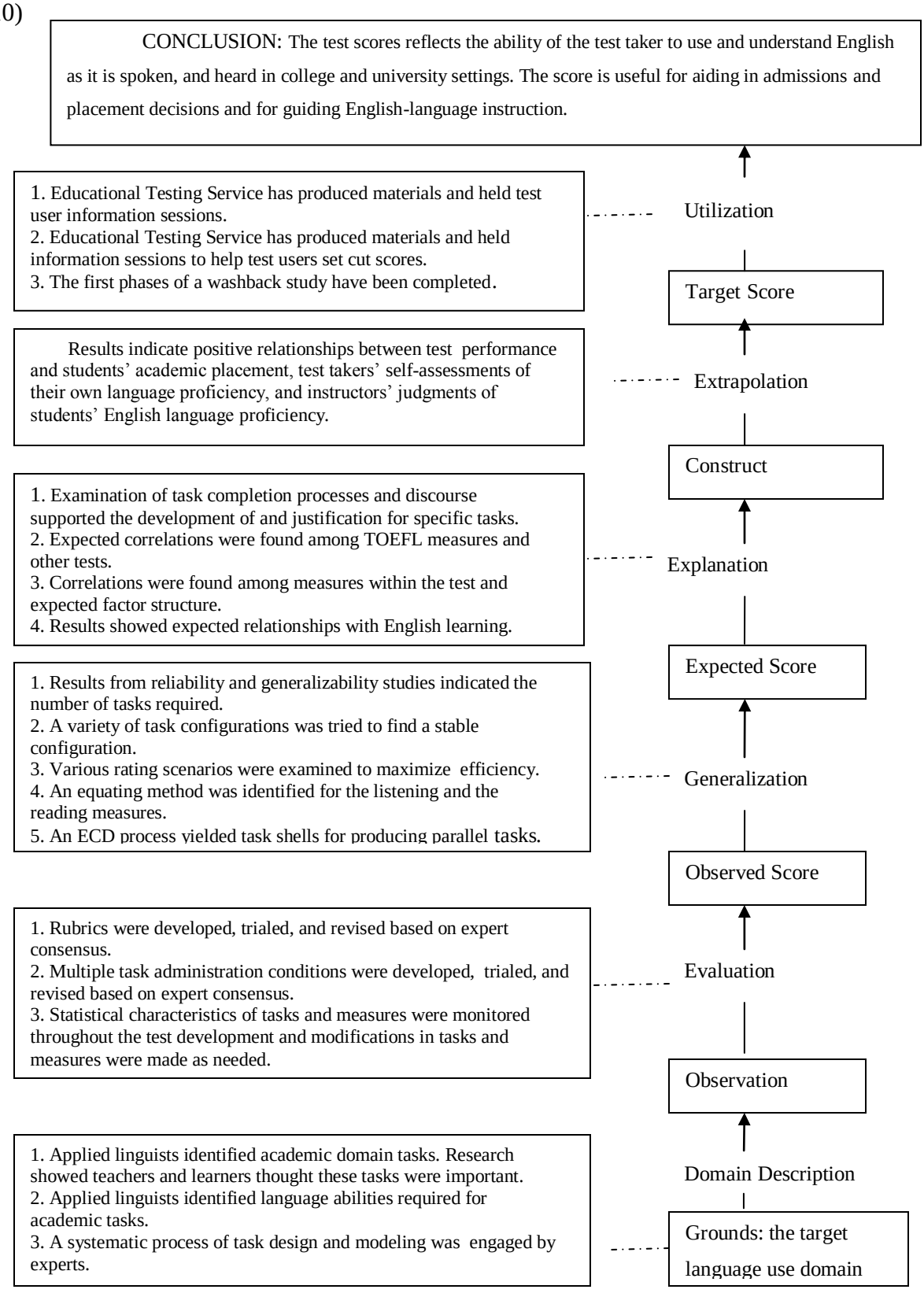
The extrapolation inference is based on the warrant that the construct of academic language proficiency measured in the TOEFL accounts for the quality of linguistic performance in English-medium institutions of higher education; in other words, performance on the test is related to other criteria of language proficiency in the academic context. Underlying this inference is the assumption that performance on the test is related to other criteria of language proficiency in academic contexts. Backing for this assumption can be found in research examining relationships of the new measures with other measures of English in an academic

context, test takers' self-assessments, instructors' judgments about students, and course placements.

Finally, the utilization inference is made on the warrant that estimates of the quality of performance in the English-medium institutions of higher education obtained from the TOEFL are useful for making decisions about admissions and comprise appropriate curricula for test takers. This inference is made on the warrant that estimates of the quality of performance in the English-medium institutions of higher education obtained from the TOEFL are useful for making decisions about admissions and comprise appropriate curricula for test takers. The assumptions for this warrant are that the meaning of test scores is clearly interpretable by admission officers, test takers, and teachers; and the test will have a positive influence on how English is taught. Some evidences supporting these assumptions can be the provision of materials or user information sessions to help users learn about the test use, or washback studies investigating testing consequences.

Based on the interpretative argument for the TOEFL, all the relevant evidences are collected and organized in order to build the validity argument for the test as being seen in Figure 3 below.

Figure 3: Structure of the validity argument for the TOEFL (Chapelle, Enright, Jamieson, 2010, p. 10)



The study acts as a model for the future argument-based validation studies in language testing and assessment. First, it structures the main components in the structure of an interpretative argument, and shows how to develop a validity argument for a high-stakes test. Another significant contribution of the construction of this validity argument for the TOEFL test is the author's suggestion that the articulation of a validity argument should consider the role of the audience as well. Accordingly, Chapelle points out the need for "differently packaged arguments for different audiences" (Chapelle, Jamieson, & Enright, 2008, p. 349).

(3) Towards a computer-delivered test of productive grammatical ability (Chapelle et al., 2010)

With the aim of supporting the potential of assessing the productive ESL grammatical ability by targeting areas identified in SLA research, and the plausibility of employing computer delivery and scoring, the researchers adopted the argument-based validation approach in order to examine the validity of a computer-delivered grammar test. The test is developed based on recent study results in SLA about the grammatical developmental path of second language learners of English with the hope of providing predictions about test-takers' grammatical ability.

The articulation of the interpretive argument as well as the outline of the validity argument for the designated test are reported to use the concepts and frameworks laid out by Kane (1992; 2001; 2006), Mislevy, Steinberg, and Almond (2003), and Bachman (2004) which are illustrated in the aforementioned validation study by Chapelle, Enright, and Jamieson (2008). Due to the fact that the test is under development without any official utilization, there are only five inferences construing the argument under examination (1) domain definition, (2) evaluation, (3) generalization, (4) explanation, and (5) extrapolation. Similar warrants and assumptions for each inference, which are presented in the earlier part about the structure of the interpretative argument for the TOEFL test, are then outlined to guide corresponding backing evidences.

Due to being a newly developed test, the focus of the validation study was finally narrowed to find theoretical and empirical evidences to support generalization, explanation, and extrapolation inferences. A number of qualitative and quantitative instruments were employed to support these three inferences. Some qualitative evidences on the test itself are the examinations of the test development and test task characteristics, scoring method, and test-taking procedures. Some other quantitative results include descriptive statistics and reliability indices,

discrimination among proficiency level groups, correlation analyses with other language tests (TOEFL, English Placement test at Iowa State University (ISU), and Writing Placement test).

(4) A comparative view about the three argument-based validation studies

The three presented studies employing the argument-based approach show the advantages of examining validity in testing and assessment as making an argument, which consequently should be judged based on its clarity, coherence, and plausibility rather than be proven. However, with the previous critical review of theoretical background of employing an interpretative argument to develop a validity argument, the examination of these three validation studies also provides an empirical evidence to support the review, which promotes the application of an interpretative argument in building a validity argument in language testing and assessment for a number of reasons. Accordingly, the framework of interpretative argument used in the last two validation studies proves to be more systematically developed than the combination of the framework of test usefulness and the concept of test purpose in the first study. In specific, instead of using descriptions of the six characteristics of test use as given by Bachman and Palmer (1996), the interpretative argument comprises of several main inferences which are well-structured with assumptions and warrants linking the test itself to the test use. These links cover all the relevant steps in the test development process. And more importantly, the interpretative argument helps to show which inferences for test interpretations should be focused, and suggests what kinds of evidences are needed to support certain assumptions in the validation study.

3. English placement test (EPT) in language testing and assessment

3.1. English placement test (EPT)

What is EPT?

Placement testing is one of the most widespread uses of tests within institutions and its scope of uses varies in situations (Brown, 1989; Douglas, 2003; Fulcher, 1997; Schmitz & C. Delmas, 1991; Wall, Clapham & Alderson, 1994; Wesche et al., 1993). Regarding its purpose, Fulcher (1997, p. 1) generalizes that “the goal of placement testing is to reduce to an absolute minimum the number of students who may face problems or even fail their academic degrees because of poor language ability or study skills”.

ESL placement testing is commonly conducted at the beginning of students’ studies to determine which level of study would be most appropriate (Brown, 1989; Douglas, 2003), and

can be put into practice in a number of ways. First, it can be used within a developmental college curricula. An example is the Written English Placement Test (WEEPT) – one of five tests in the Comparative Guidance Program (CGP) published by the College Entrance Examination Board, which was developed specifically as a guidance and placement tool for 2-year college students in order to place students in either remedial-level courses or a college-level composition course (Schmitz, & delMas, 1991). Second, it can be used for placement of students of varying language backgrounds and skill levels in an intensive ESL program (Wesche et al., 1993). In another case, a placement test can be developed to identify overseas students entering an English-medium university whose language skills or abilities are insufficient for their academic life (Douglas, 2003; Fulcher, 1997). In fact, besides using one of the major international tests such as TOEFL, or IELTS for admissions, many colleges and universities do some further evaluation of students after their arrival on campus in order to get a more precise assessment of the specific English language abilities of students. The test results will be used to decide whether the test-takers need more English instructions or not, and which appropriate ESL courses can be offered to meet their needs (Douglas, 2003, p. 4).

Brown (1996) presents some further descriptions about EPT. First, program-level EPT tests aiming at grouping students into similar ability levels are usually norm-referenced (Brown, 1996, p. 21). Accordingly, a norm-referenced test (NRT) is designed to measure global language abilities, and each student's score on such a test is interpreted relative to the scores of all other students who take the test. The score results of a norm-referenced test or an EPT are thus expected to spread out as a bell curve. Next, EPT tests have some differences from proficiency tests (Brown, 1996, p. 11). While a proficiency test tends to be very general in character, because it is designed to assess extremely wide bands of abilities. A placement test must be more specifically related to a given program, particularly in terms of the relatively narrow range of abilities assessed and the content of the curriculum, so that it efficiently separates the students into level groupings within that program. Hence, a general proficiency test might be useful for determining which language program is the most appropriate for a student; once in that program, a placement test would be necessary to determine the level of study from which the student would most benefit.

What are the impacts of EPT?

Based on the potential impacts of decisions made on test-takers' performances on a test, placement testing is considered to be a mid-stakes test on the scale from low-stakes to high-stakes (Bachman, 1990, 2004; Douglas, 2003). While high-stakes decisions are major, life affecting and its wrong decisions cause high costs, effects of low-stakes decisions are relatively minor with much lower possible costs caused by wrong decisions (Bachman, 1990). The middle-range impacts of placement decisions can be explained here. First, the reliability and validity of its decisions in sorting students into relatively similar ability groups have influences on the effectiveness of language programs (Brown, 1989; Fulcher, 1997; Schmitz & C. Delmas, 1991; Wall, Clapham & Alderson, 1994; Wesche et al., 1993). For example, the accurate and consistent placement of the students into their language proficiency helps language instructors responsibly serve their needs, and manage the content to teach. Second, placement decisions can also affect the lives of the students involved in terms of the amounts of time, money, and effort that they will have to invest in learning the language. For instance, it will cost time and money or cause emotional impacts such as frustration if a student is mistakenly placed in a wrong class where his proficiency is too much lower or too much higher than those of other peers. In brief, the decisions made on a placement test generally do not substantively affect its test-takers' lives, as well as other test users; however, wrong decisions are possible to affect its test-takers or other users in terms of finance, time, or emotional impacts.

What are major issues of EPT in practice?

Brown (1996, Chapter 2) discusses several theoretical and practical issues the interactions of which have great influences on the decision of adopting, developing or adapting a language test for placement in language programs. Some theoretical issues are how to define and describe the language framework, or how to balance the relationships among competence, performance, and test tasks for a placement test. In fact, Douglas (2003, p. 4) highlights this issue as examining how to develop on-campus English language placement testing in colleges and universities. According to him, the relationship between language knowledge and content knowledge in specific academic field still poses a major issue in the assessment of academic English in general, and in any on-campus testing context. He says that "the more specific the purpose of the test, ranging from a general academic writing test to a quite specific test of business report writing, the more the specific content knowledge gets entangled with language

knowledge" (Douglas, 2003, p. 4). This issue leads to a concern for the interpretation of test results because it is difficult to decide the proportions of content knowledge and language knowledge in test-takers' performances, and their test scores. For instance, test-takers can mainly base on their academic background instead of their language knowledge and skills to do the test if the test involves more specific content knowledge. Other practical constraints upon EPT testing are fairness, cost, and other logistical issues such as ease of test construction, ease of test administration, and ease of test scoring. These concerns are finally suggested to be taken into consideration as judging validity of an EPT in a language program.

3.2. Validation of an EPT

A number of researchers have been trying to address the issue of validity in placement testing (Brown, 1989; Fulcher, 1997; Lee, & Greene, 2007; Schmitz & DelMas, 1991; Truman, 1992; Usaha, 1997; Wall, Clapham, & Alderson, 1994). Some major concerns and approaches in examining validity of an EPT can be summarized here.

(1) Issues in reliability and validity of a placement test (Fulcher, 1997)

Despite the popularity in use within institutions, there is relatively little research literature relating to reliability and validity of language placement tests (Fulcher, 1997; Schmitz, & DelMas, 1991; Wall, Clapham, & Alderson, 1994). Also, most of validation studies of EPT tests adopt the earlier accumulation-of-evidence validation approach in which different types of validity are examined separately for such a test (Fulcher, 1997; Schmitz & DelMas, 1991; Wall, Clapham, & Alderson, 1994). For instance, in the pioneering validation study of an English placement test designed to screen students entering a British university, Wall, Clapham, and Alderson (1994) provided a collection of evidences including face validity (student perceptions of the test), content validity (tutors' evaluation on the representativeness of test content in comparison with program content), construct validity (how significant the correlations in performances among different tests), concurrent validity, and reliability statistics. Taking this approach, Fulcher (1997) continued to investigate validity of the language placement test at the University of Surrey the purpose of which is to identify students needing more English instructions to be successful in their academic life. The test is about one-hour long, and consists of three parts: (1) Essay writing, (2) Structure of English, and (3) Reading Comprehension. In order to provide evidences on the validity of the test, besides using the methods in the study by Wall, Clapham, and Alderson (1994), he also elaborated other aspects of the test including how

to set cut-scores for placement, how to exploit more means of statistical analyses, how to develop parallel test forms, and how to use student questionnaires for face validity.

Another significant issue in validating an EPT test is how to take into account a number of relevant constraints as examining validity of an EPT test (Fulcher, 1997). In specific, these factors comprises of economical, logistic and administrational constraints. For example, it can be how much testing time allowed, or how many examiners employed, or how much money and efforts spent on carrying out pretesting and post hoc analyses, or equating test forms and formats.

(2) Typical inferences in an interpretative argument for EPT

Based on Messick's (1989) theoretical work in construct validity, Schmitz and delMas (1991) have made a great contribution to how to validate a placement test by exemplifying how to examine major inferences in EPT test score interpretation and use. First, they state two main inferences in placement test score interpretation and use which are followed by their clarification of the underlying hypotheses of these inferences. Then, they offer some guidelines on validating placement decisions. Finally, they illustrate how to use these hypotheses and guidelines through a validation study of the Written English Placement Test (WEPT).

Placement tests are described to share two most common inferences in interpretation and use which should be identified in their validation studies (Schmitz & delMas, 1991, p. 31). First, scores accurately represent a student's standing within an academic domain or dimension of learning. The second is that a certain amount of mastery within that domain is required for the student to succeed in a college-level course or curriculum. These two inferences thus reflect the essential role of placement tests which is to discriminate among students who need to take remedial-level work from those who do not, or among those who need different levels of instructions.

Next, these two main inferences are elaborated to comprise of four possible underlying hypotheses that should be considered in validating placement tests (Schmitz & delMas, 1991, p. 40).

1. The test distinguishes between masters and non-masters within an academic domain of learning.
2. Placement scores contribute to the prediction of course grades in sections for which student placement was unguided by test scores.
3. Placement of students according to placement test cut scores results in higher rates of course success (hit rates) than rates achieved when placement scores are not used (base rates).

4. Course success is related to other criteria representing desirable standards, for example, performance in subsequent courses and cumulative grade point average (GPA).

These four hypotheses in the investigation of validity of a placement test are claimed to rest on four main assumptions. First, courses in the local curriculum are built on a hierarchical sequence of concepts or skills, and that mastery of foundation concepts or skills in lower courses, is, in fact, necessary for success in higher courses. Second, incoming students differ from each other with respect to the dimension being assessed. Third, student performance in the course or curriculum shows variation. Because the purpose of the test which is to distinguish between masters and nonmasters in a domain, test-takers are expected to show variation in their test scores. A fourth assumption concerns the relation of the test content to the curriculum content. Accordingly, a valid placement test for developmental courses is one in which the skills and concepts being assessed in the test are similar in nature to those that are taught and assessed in the curriculum.

Based on the four suggested underlying hypotheses for using a placement test, the authors continue to give some guidelines on how to examine different validity types of a placement test (Schmitz & delMas, 1991, p. 41-42). First, correlations between placement scores and course grades can be useful in giving evidences on predictive validity. Also, such evidence, which proves how the placement test contributes to the prediction of course grades, is recommended to be gathered in the applied setting before making placement recommendations. Regarding validating cut-scores, the authors suggest using decision-theoretic approaches to support the plausibility of decisions made in placement tests. Finally, the validation of a placement test use should include an investigation into benefits gained from using placement scores. For example, based on the third hypothesis, hit rates can be calculated on courses after using placement systems, and be compared against the baseline data.

In brief, the review of validation studies of placement testing in general, and EPT in particular is meaningful in a number of ways. It reveals current issues in how to examine validity of an EPT, and emphasizes the need to figure out a way to judge or evaluate a placement test use effectively. Second, some main constituents of validity for a placement test use or interpretation are suggested. Moreover, few validation studies have specifically dealt with on-campus placement testing for admitted international students into English medium higher-education

programs (Fulcher, 1997; Usaha, 1997; Wall, Clampham, & Alderson, 1994). These facts motivate me to use the current argument-based approach in order to address the current issues in validity of an EPT, especially the specific EPT testing for new international students coming to English-medium colleges or universities.

3.3. Testing and assessment of listening in second language

Listening comprehension in second language

Considered as one of the essential components of any language test and language skills in communication, listening skills in English as a second language have attracted a lot of researchers' attention in order to understand the listening comprehension process thoroughly, and to develop an effective listening test (Buck, 2001; Feyten, 1991; Rost, 2002). Listening comprehension is described as a very complex process involving both linguistic knowledge such as phonology, lexis, syntax, semantics, discourse structure and non-linguistic knowledge. Also, the processing of the different types of knowledge in the listening comprehension process is agreed to not occur in a fixed order. Thus, listening comprehension is the result of an interaction between a number of information sources, and listeners have to employ their skills and knowledge to decode and interpret what is heard.

The testing and assessment of listening skills in second language plays an important role for a number of uses (Buck, 2001; Rost, 2002). First, it partly contributes to the overall assessment of a second language learner's language ability. In fact, listening skills are suggested to construe a tremendously important aspect of overall language ability because more than 45% of our total time communicating is spent listening (Feyten, 1991). Second, in some testing situations with limited resources, listening is claimed to be able to substitute for other oral skills, i.e. speaking skills due to their high correlation in performances (Buck, 2001, p. 96). In other words, test-takers' results on listening tests can be used to give some information on their speaking performances. Next, listening tests can be made for assessing achievement, admissions, placement and diagnosis specifically in language acquisition.

Academic listening in second language

Second-language listening in an academic setting has tremendously intrigued testing researchers due to its importance in academic life (Chiang & Dunkel, 1992; Flowerdew, 1994; Hanson & Jensen, 1994; Jensen & Hanson, 1995). Academic listening usually involves listening to lectures or presentations on academic topics in a college or university. Some characteristics of

academic listening include its primarily non-interactional nature, the need for skills in handling specialist vocabulary and long stretches of speech (Flowerdew, 1994), the sub-skills required to decode an academic lecture such as note-taking , the audio-visual aspect of academic listening, and the place of authentic listening texts and activities in the teaching of academic listening. Indeed, some sub-skills such as note-taking, inferences, or guessing are crucial for non-native speakers in the academic setting, for whom sound system problems appear to present particular challenges (Brown, 1990), and vocabulary problems are proved to be a significant barrier to listening comprehension for advanced learners (Kelly, 1991). Noticeably, Richards (1983) gives a framework of sub-skills in academic listening which helps to demonstrate how demanding and complex the process of academic listening comprehension is (see Table 2 below).

Table 2: A framework of sub-skills in academic listening (Richards, 1983)

1. Ability to identify purpose and scope of lecture
2. Ability to identify topic of lecture and follow topic development
3. Ability to identify relationships among units within discourse (e.g. major ideas, generalizations, hypotheses, supporting ideas, examples)
4. Ability to identify the role of discourse markers in signaling structure of a lecture (e.g. conjunctions, adverbs, gambits, routines)
5. Ability to infer relationships (e.g. cause, effect, conclusion)
6. Ability to recognize key lexical items related to subject/topic
7. Ability to deduce meanings of words from context
8. Ability to recognize markers of cohesion
9. Ability to recognize functions of intonation to signal information structure (e.g. pitch, volume, pace, key)
10. Ability to detect attitude of speaker toward subject matter
11. Ability to follow different modes of lecturing: spoken, audio, audio-visual
12. Ability to follow lecture despite differences in accent and speed
13. Familiarity with different styles of lecturing: formal, conversational, read, unplanned
14. Familiarity with different registers: written versus colloquial
15. Ability to recognize relevant matter: jokes, digressions, meanderings
16. Ability to recognize function of non-verbal cues as markers of emphasis and attitude
17. Knowledge of classroom conventions (e.g. turn-taking, clarification requests)
18. Ability to recognize instructional/learner tasks (e.g. warnings, suggestions, recommendations, advice, instructions)

Constraints in testing listening in second language

There are some theoretical and practical issues in how to measure listening abilities in second language accurately and effectively. The significant theoretical issue in second language listening testing and assessment is how to define listening constructs. First, there are a number of unique characteristics in second-language listening comprehension, which cause difficulties to English-listening test developers (Buck, 2001). As emphasizing the differences between second language listening and first language listening, Buck (2001, p. 49) points out that second-language listening comprehension is more conscious, and requires the use of compensatory skills. Moreover, the use of these sub-listening skills varies in accordance with situations or purposes leading to the complex nature of the listening process (Richard, 1983). For example, listening can be categorized into conversational listening, listening for entertainment, and academic listening. Second, Buck (2001, p. 32-39) gives a number of features influencing listening comprehension that challenges the design and development of a listening test. They include phonology, accent, prosodic features, speech rate, hesitations, and discourse structure.

Several practical constraints in developing a listening test can be summarized here (Buck, 2001). First, due to high costs, most of listening tests are non-collaborative listening. In other words, these tests measure test-takers' listening abilities by their understanding of what speakers mean in non-interactive situations. Second, varying interpretations from listening texts and limitations on providing channels in listening tests create big challenges in defining listening construct, and designing listening test items or tasks. The argument for this is that effective real-life communication does not always require a total and precise understanding through listening, but relies on other factors such as cooperation, and inference. In addition, the way to test test-takers' listening comprehension is through other channels which require other abilities irrelevant to listening comprehension such as reading or writing. For example, it is necessary to read given options to complete a multiple-choice listening test, or to have good working memories to succeed in listening to lectures. In another way, open-ended questions that require the test-taker to construct a response would require less reading, or less memorization, but then writing is needed. The last but not the least, available resources including qualified test developers, material resources, and sufficient time have a great impact on how to develop a good listening test (Bachman & Palmer, 1996).

4. Summary

Based on the above review of current validation studies in language testing and assessment, especially EPT in colleges and universities, I would like to investigate the validity of the Listening EPT test used at Iowa State University (ISU), which is administered to international new comers whose first language is not English. With the aim of addressing the current issues in validation studies of EPT – a mid-stakes testing, the current argument-based approach is adopted in order to build up a validity argument for the Listening EPT test at ISU. In specific, based on the structure of interpretative argument, and validity argument explored in the study by Chapelle, Enright, and Jamieson (2008, 2010), as well as some suggested hypotheses and inferences in using placement tests, the interpretative argument for the Listening EPT test at ISU will be structured. However, due to time constraint, not all the inferences of the interpretative argument for the EPT test can be examined in my study; instead, only some main inferences will be investigated.

Using the framework of the interpretative argument for the TOEFL test developed by Chapelle, Enright, and Jamieson (2008), I propose the interpretative argument for the Listening EPT test consisting of six main inferences: Domain description, Evaluation or Observation, Generalization, Explanation, Extrapolation, Utilization. Accordingly, each inference comprises of corresponding assumptions and warrants which help to structure the validity argument for the Listening EPT test. Table 3 below presents the suggested construction of the validity argument developed for the Listening EPT test which is purposed for future investigation as well. In the study, four out of six inferences (Warrants 1, 2, 3, 4) will be studied. In specific, the four research questions under examination in the study are:

1. How do the EPT Listening test' design and development help to measure what we want to measure of test-takers? (Warrant 1 & 2)
2. How reliable is the EPT Listening test in measuring test-takers' proficiencies? (Warrant 3)
3. How do students' scores on other test of language development (TOEFL) correlate with their scores on the EPT Listening test? (Warrant 4)
4. What are challenges to the validity argument of the EPT test at ISU?

Table 3: Summary of the inferences, warrants in the validity argument with their underlying assumptions for the EPT listening test at ISU (based on the TOEFL validity argument given by Chapelle, Enright, Jamieson (2010, p. 7)

Inference	**Warrant Licensing the Inference**	**Assumptions Underlying Inferences**
Domain description	**Warrant 1:** Observations of performance on the EPT Listening test reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education, especially in Mid-western areas of the U.S.A.	1. Critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified. 2. Assessment tasks that require important listening sub-skills and are representative of the academic domain can be simulated.
Evaluation	**Warrant 2:** Observations of performance on EPT listening tasks are evaluated to provide observed scores reflective of targeted language abilities (academic listening proficiency).	1. Rubrics for scoring responses are appropriate for providing evidence of targeted listening abilities. 2. Task administration conditions are appropriate for providing evidence of targeted listening abilities. 3. The statistical characteristics of listening test items, measures, and test forms are appropriate for norm-referenced decisions.
Generalization	**Warrant 3:** Observed EPT listening scores are estimates of expected scores over the relevant parallel versions of listening tasks, test forms, and across raters.	1. A sufficient number of tasks are included on the EPT listening test to provide stable estimates of test takers' listening performances. 2. Configuration of tasks on listening measure is appropriate for intended interpretation. 3. Appropriate scaling and equating procedures for EPT listening test scores are used. 4. EPT listening task and test specifications are well defined so that parallel tasks and test forms are created.
Explanation	**Warrant 4:** Expected listening scores in the EPT Listening test are attributed to a construct of academic listening proficiency.	1. The linguistic knowledge, processes, and strategies required to successfully complete listening tasks vary across tasks in keeping with theoretical expectations. 2. Task difficulty is systematically influenced by task characteristics. 3. Performance on the EPT listening test relates to performance on other test-based measures of language proficiency as expected theoretically. 4. The internal structure of EPT listening test scores is consistent with a theoretical view of language proficiency as a number of highly

Inference	Warrant Licensing the Inference	Assumptions Underlying Inferences
		interrelated components. 5. Test performance on the EPT Listening test varies according to amount and quality of experience in learning English.
Extrapolation	**Warrant 5:** The construct of academic listening proficiency as assessed by the EPT accounts for the quality of linguistic performance, especially listening performance for academic purposes in English-medium institutions of higher education.	Performance on the EPT Listening test is related to other criteria of language proficiency in the academic context.
Utilization	**Warrant 6:** Estimates of the quality of performance at ISU obtained from the EPT Listening test are useful for making decisions about appropriate curricula for test takers, and successful communication in academic life.	1. The meaning of EPT Listening test scores is clearly interpretable by department officers, test takers, teachers and other relevant parties. 2. The EPT Listening test will have a positive influence on how students should prepare their academic listening proficiency at ISU.

CHAPTER 3: METHODOLOGY

This chapter consists of two main parts. The first part is aimed at providing an overview of the context of the study, specifically the English Placement Test (EPT) at Iowa State University (ISU), with a detailed description of the EPT Listening test used in Fall 2010. The second part focuses on my rational selection of the methodology for my study. It presents three instruments, and how to implement each of them in the study.

1. Context of the study

1.1. Description of the EPT test at ISU

1.1.1. About the test

EPT Test history

Iowa State University (ISU) has employed the EPT test to examine the language proficiency of new international students for a long time. The test is under the authority of the English Department at ISU, and has been managed by a number of personnel who are professors at the English Department. However, the history record of the test consisting of test booklets and test result data did not start until the summer of 2007 under the supervision of Prof. Volker Hegelheimer.

Due to the unavailability of information about the whole EPT test history, a brief overview of the EPT test history from Summer 2007 to Fall 2010 is given here. During this period, about 11 examinations were administered to more than 2,000 test takers, and five sets of test booklets were written for use (Set Summer 07, Set A, Set B, Set C1, and Set C2). Also, a number of revisions on the EPT test booklets have been made based on reliability estimates and test item analyses. For example, the number of test items in an EPT test is finalized to be 30 to ensure a sufficient reliability estimate under other practical time constraints of running the EPT test. The two test sets in 2007 (Set Summer 07 and Set A) have the largest number of test items (38 and 40 items). Set C1 used in Spring 2010 has only 25 items while Sets B and C2 both have 30 items which is proved to be more appropriate for a one-hour long test. Another significance among these different test booklets is the development. Set A was mainly developed from Set Summer 07 with the addition of some new items, and the replacement of some low discriminate items. Moreover, most of the items in these two sets were reported to run a pilot test before use. Set B was developed based on these two test sets and included some more additional items

whose quality had not been attested before use. The whole new C1 set was developed without any pilot tests. Set C2 was based on Set C1 with some changes.

Table 4: Test Booklet History from Summer 2007 to Fall 2010

Set	Set Summer 07	Set A	Set B	Set C1	Set C2
Semester	Summer 07	Fall 07	Spring 08 Summer 08 Fall 08 Spring 09 Summer 09 Fall 09	Spring 10	Summer 10 Fall 10

EPT Test takers

As stated in the EPT administration and result processing manual created as of May, 2010, "the EPT is administered to non-native English speaking students who enter Iowa State University as they are required to meet the English requirement either by passing the placement test or by completing required ESL courses unless they meet some exemption categories." (EPT administration and result processing: manual, 2010). Accordingly, there are two main exemption categories : (1) one for those whose scores on some internationally standardized tests exceed a certain test score requirement (see Table 4 below for details), and (2) one for those who are from some countries where English is the primary or official language. In other words, the test is designed for both admitted ISU undergraduate and graduate students whose native languages are not English. Its purpose is to check whether their English proficiencies are sufficient for their studies at ISU, or whether they need more English instruction or not. In case of needing more English instruction, they will be placed into supplementary language classes of different language skills and levels based on their scores, and required to complete them before graduation.

Table 5: Non-native English speaking students exempt from the English Placement Test at ISU

Group 1: Non-native English speaking ISU students who meet or exceed any of the following test scores: TOEFL – 105 or higher (iBT); 640 or higher (PBT) IELTS – 8.0 or higher ACT (English) – 24 or higher SAT (Verbal) – 550 or higher *Group 2:* Students who already have a bachelor's, master's, or Ph.D degree received from a university where English is the only language of instruction, or an accredited four-year college or university within the U.S.

EPT Test structure, testing method and scoring rubrics

The EPT test comprises of three sections, essay writing, listening and reading comprehension, which takes approximately 3 hours in total. The test starts with a one-hour long writing section, followed by reading and listening sections after a 10 minute break. Each reading and listening section is estimated to take 40 minutes to complete. Scores on each component will be used to assess each individual language skill of test-takers separately.

The EPT Reading and Listening tests are paper-based, and employ the multiple-choice format for time-efficiency. Only one partially constructed response is used in the EPT Reading and Listening tests. In specific, each question in the EPT Reading and Listening tests provides four choices, and asks students to choose the best answer. Questions can fall into the categories of inference or comprehension checking. Both tests have a strong emphasis on English for academic purposes. In the reading section, there are about three to five academic passages of about 600 words, each of which is followed by 10 to 12 questions. The Listening section comprises of four lectures with about 30 questions in total. The test-takers record their answers on computer forms which are sent to the Solution Center at ISU for automatic scanning and scoring. Their EPT Listening and Reading scores are counted based on the number of correct answers without any difference in weighting among the questions.

Test administration

Students are supposed to take the EPT test upon their arrival to campus because the test results will be used to decide whether they need more English instruction for their studies at ISU

or not. The test is held every semester. In fall semesters, there are three regular test sessions given before or during the orientation week – one on Friday immediately before the orientation week and the other two on Monday afternoon and Tuesday morning during the orientation week. In spring and summer semesters, only one regular test session is provided for students. A make-up test is also administered to late arrival students in spring and fall semesters, and is on Tuesday evening of the first week of the semester.

The administration and result processing of the EPT test are carefully described and instructed in the EPT test manual created in May, 2010. Accordingly, an EPT test administration involves a number of different tasks such as preparing a testing environment (reserving rooms with a sufficient number of seats, and necessary equipment, preparing test materials, finding proctors, and listing test-takers' information, ect.), and giving the test (checking students, sorting the record sheets, and enrolling students in the EPT WebCT course, ect.). For result processing, the final test result of each test-taker will be recorded on three different sheets delivered to relevant parties (the Graduate College, the EPT office, and students). Noticeably, score reporting requires ample efforts due to time pressure as test results are supposed to be available within one day after the test.

1.1.2. Test purpose

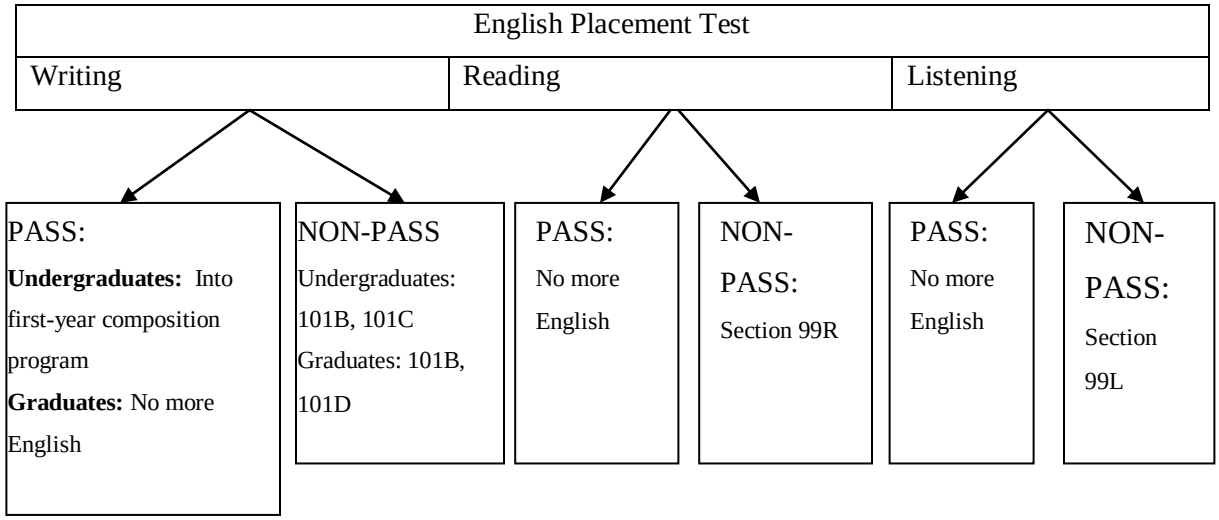
The test purpose of the EPT test at ISU will be described based on the concept given by Stoynoff and Chapelle (2005). According to them, test purpose can be elaborated in terms of "three dimensions that capture the important functions of the test" which include inferences made from the test, the uses of the test and the scope of the impact of the tests (Stoynoff & Chapelle, 2005, p. 10).

The first dimension concerning the inferences drawn from the test scores, is described on a continuum that ranges from specific (where connections are made to what is explicitly taught) to general (where the test measures general language ability). Based on the test description stated in the EPT information sheet for orientation purpose, "the English Placement Test is designed to test students' academic writing, reading and listening ability." (Appendix F, EPT administration and result processing: manual, 2010). Thus, inferences about the test takers language ability for the ISU English Placement Test fall more towards the general side of the continuum, showing the test taker's academic English language proficiency.

The second dimension, which includes the educational uses or decisions made on the basis of test results fall on a continuum that ranges from low to high stakes. As introduced in the EPT manual, "the English Placement Test, which is given at the beginning of the spring, summer and fall semesters, is to determine whether ISU students whose native language is NOT English are proficient enough in English to meet requirements at Iowa State University" (Appendix F, EPT administration and result processing: manual, 2010). More specific decisions based on EPT test scores can be given here. First, graduate students who pass the test meet the Graduate College requirement for certification in English, and do not have to take any English courses unless they are required to do so by their departments. And undergraduate students who pass the test are eligible to take English 150 – a course required of all undergraduate students regardless of native language. On the other hand, all the students who do not pass the test will be required to take one or more English supplementary courses.

More importantly, the test-takers' EPT score result on each component of the EPT test will be used to place them into different English courses of different skills and levels. They are advised to enroll in English supplementary courses within the first year while taking other courses in their academic programs, and fulfilling their English language requirements is a condition for graduation. There are five supplementary English courses offered for placement: (1) 101B (Academic English 1 for graduates and undergraduates) focusing on a review of English grammar in the context of writing and basic English academic writing at paragraph level; (2) 101C (Academic English 2 for undergraduates) preparing students with techniques of English academic writing at essay level; (3) 101D (Academic English 2 for graduates) instructing how to write professional communication, academic papers, and reports; (4 & 5) 99L & 99R (Academic Listening and/or Reading for graduates and undergraduates) concentrating on improving academic listening and reading skills respectively. Placement decisions can be illustrated in the diagram below (see Figure 4).

Figure 4: Placement for non-native speakers of English at Iowa State University (ISU)



So in comparison with the TOEFL iBT whose scores are used for a high-stakes decision, i.e. admissions to university, the ISU English Placement Test falls towards the middle of the continuum as the placement decisions do not affect the lives of test-takers despite the fact that their wrong decisions will cause costs to the relevant test users, and influence the effectiveness of the supplementary English language program (Brown, 1996; Douglas, 1998).

Next, the third dimension concerns the scope of impact of language tests on relevant parties and activities such as test-takers, teachers, the society, or language teaching and learning activities. Based on the purpose and the major placement decisions of the EPT test results, the ISU English Placement Test might have less of an impact on society, but it still has a broad impact overall, varying from an impact on students to other test users including teachers of English classes, test-takers' advisors, and departments. For example, for students, the test results have an impact on their study plans and budget for paying required English courses, and they will focus more on equipping with English language knowledge and skills for academic purposes. Besides, for instructors of English classes, the placement of homogeneous students into a certain language class of a certain language level is of importance in order to orientate and achieve course objectives. Likewise, the EPT test results are expected to give some important information about students' language proficiency contributing to their advisors' consultancy.

1.2. Description of the EPT Listening test – Fall 2010 at ISU

1.2.1. Test purpose

As described in the test purpose of the EPT test at ISU above, the EPT Listening test – a component of the EPT test at ISU has the following purpose. The test is developed in order to measure ISU international students' academic listening abilities. Its results will be used to make inferences about test-takers' academic listening abilities, for example, their proficiencies in listening to lectures. Such inferences from their EPT Listening test will be used to make a placement decision that is whether test-takers need more English instructions in academic listening or not in order to be successful in their academic life at ISU. In other words, their scores will be used to decide if they have to take the supplementary English course 99L whose objective is to provide instructions on strategies and techniques in improving English listening skills for academic purposes. The students' performances on the EPT Listening test at ISU can also be beneficial for advisors who want more information and evidence on their students' academic English language proficiencies.

In terms of test impact, as presented in the test purpose of the EPT test at ISU, the EPT Listening test is expected to have impacts on its test-users, especially students. For students, whether they pass the test or not will interfere their study plan, and financial budget. For instructors of the Listening class 99 Section R, the reliability and validity of the inferences based on the test results and the placement decision will influence the effectiveness of their supplementary instructions. For example, if the placement decision is not plausible and assigns test-takers in a wrong class, instructors will take more time and efforts in reassessing proficiencies of students in the class, and find it more challenging to deal with a class which has students possessing a wide range of listening proficiencies.

1.2.2. Administration of the EPT Listening test – Fall 2010 at ISU

A brief report on how the EPT Listening test happened in Fall 2010 can be given here. Included in the report is information about date, time, location of the test, and how the test was operated. Noticeably, some observations on the administration of the test are also provided.

General information

The main person in charge of administrating the EPT Listening test in Fall 2010 was Yoo-Ree Chung under the supervision of the test coordinator - Prof. Volker Hegelheimer. The

EPT Listening test in Fall 2010 comprised of three regular tests, and one make-up test which was for late arrivals to ISU. The dates and times of these tests can be found below (see Table 5).

All the three tests in Fall 2010 took place in Room 125 in the Kidlee Hall. The room is large and well equipped with a good auditorium system for listening, and two large screens and projectors for showing videos, and giving instructions; however, for the very back rows, it is sometimes difficult for the back rows to see videos clearly, especially some subtitles in the Listening test. Each EPT test administration took about nearly four hours to complete all the major tasks from checking students to the final step of collecting all the required papers.

Table 6: Summary of the EPT Administration for Fall 2010

Test Date	Time	Undergraduates	Graduates	Students
8/14/2010	12pm-5pm (The test begins at 1pm)	329	163	95
8/16/2010	12pm-5pm (The test begins at 1pm)			217
8/17/2010	9am-2pm (The test begins at 10am)			180
8/24/2010	5pm-10pm (The test begins at 6pm)	48	17	65
TOTAL		**377**	**180**	**557**

Test-takers

The total number of test-takers of the EPT Listening test in Fall 2010 was 557. Nearly 68% of the EPT test takers in Fall 2010 were undergraduate students, and the rest were graduate students. While the first three tests had the majority of test-takers, there were around 65 students in the make-up test, which was observed to be convenient and much easier to administer the test. All the test-takers were informed of taking the test as receiving their admissions from the Graduate College, and were provided with instructions on how to register and prepare for the test which could be announced during the Orientation week, or be found on the website *http://www.grad-college.iastate.edu/about/englishexam.html*.

Test score sets

As reported above, 557 students participated into the EPT Listening Fall 2010 administration; however, the score set of the EPT Listening Fall 2010 administration is comprised of 556. Also, 395 of these EPT test-takers in Fall 2010 had their scores on the internationally English language standardized tests developed by the Educational Testing Service (ETS) available. The test is called Test of English as a foreign language (TOEFL). It also offers

two versions known as TOEFL iBT and TOEFL pBT which are based on two different manners of test delivery, namely Internet-based and paper respectively. The TOEFL score data set of 395 EPT test-takers in Fall 2010 consists of 344 TOEFL iBT scores and 51 TOEFL pBT scores. For better comparison, these 51 TOEFL pBT scores were converted into equivalent TOEFL iBT scores using the conversion chart published by the ETS (2008).

In addition, the majority of the TOEFL iBT scores of these test-takers include the listening component scores. Accordingly, 268 out of 344 EPT test-takers in Fall 2010 who reported the TOEFL iBT scores, had the TOEFL iBT Listening component score available.

Placement decisions based on the EPT Listening scores

In general, the placement decision is made based on the test-takers' scores on the EPT test. So does the EPT Listening test. EPT test administrators set a cut-off score which is used to decide whether a test-taker passes or fails the EPT Listening test, or whether he or she has to or does not have to take the 99L course. The cut-off score is reported to be preset, which was 13 out of 30 for the EPT Listening Fall 2010 administration at ISU. Accordingly, the cut-off score was determined by using a few criteria: (a) descriptive statistics (especially, mean and median), (b) a 40/60 rule of thumbs, and (c) the availability of listening sections (and instructors). The test administrators explained the 40/60 rule of thumb as following. They used to pass students who got 60% or more of the listening items right in the old EPT test. As the difficulty level of the test increased quite a bit through several revision sessions, they had to lower the cutoff score, and ended up passing students who got 40% or more of the listening items right eventually. For the test set used in the EPT Fall 2010 administration (Set C2), they also considered the mean and median of the collected test scores at its first administration, which were around 13. In addition, the availability of the ESL courses to be offered was also taken into consideration in the course of decision- making. Thus, a few different scenarios with slightly different cutoff scores (e.g., 12, 13, and 14) were created with the counting of the number of potential passes in each scenario which lead to the final cutoff score for the EPT Listening Fall 2010 administration of 13 in the end.

Table 6 below presents a summary of placement decisions based on the available data of the test-takers of the EPT Listening Fall 2010 administration at ISU. The placement results are categorized into groups of the EPT Fall 2010 Listening test-takers based on their reported TOEFL scores for admission.

Table 7: Summary of placement decision results of the EPT Listening Fall 2010 test takers at ISU in correspondence with different score sets.

Test	**Count**	**Placement decision 99L (Fail)**	**Placement decision (Pass)**
TOEFL pBT	51	19	32
TOEFL iBT Listening	258	50	208
TOEFL iBT total score	344	62	282
TOEFL iBT total score (with converted TOEFL pBT score)	395	81	314

Test booklet – Set C2

The test booklet used for the EPT test in Fall 2010 is Set C2. The set (C2) comprises of two main sections: Reading and Listening. The EPT Listening test in Fall 2010 (set C2) adopted the multiple - choice format, and all the 30 questions with instructions were printed in the test booklet. Each of the test-takers was given a test-booklet and was instructed on how to proceed the test booklet by the proctor through the loudspeaker.

As described in the EPT test history above, Set C2 was almost based on Set C1 which was used in Spring 2010 with one additional lecture. Accordingly, the first three out of the four lectures in Set C2 (or Set C1) were developed by the students as a required assignment in the course of language testing and assessment taught by Prof. Dan Douglas. These lectures with listening questions were reported to be submitted on 30 April 2009. These lectures were developed by following a number of steps. The test developers were informed about the EPT test purpose, and the test characteristics in order to find appropriate materials and design suitable test tasks or questions. The developed questions with the three lectures finally underwent a test pilot with the participation of eight to ten students. Based on the test pilot results, bad items were revised or removed leading to the finalized EPT listening test set (Set C1) or a major part of Set C2. The fourth lecture in Set C2 was created by Yoo-Ree Chung, and was reported to be reexamined by Prof. Volker Hegelheimer for its content, and appropriateness of its questions. A closer examination of the test booklet will be provided in the next chapter.

Other resources

Some main tasks in administering the EPT Listening at ISU in Fall 2010 were checking in the test-takers, arranging their seats, handing and collecting relevant papers and test booklets,

and preparing test score report papers. Each task involved a number of resources including human, materials and equipments.

Check-in for test-takers

The check-in process includes checking student name, ID number, and ID net account. Some cases did not have an ISU card, and a net ID yet, so the only way was to check their passport to check their names, and photos. Some instructions on how to do check-in for students are also presented in the EPT test administration and result processing manual (2010). After check-in, test-takers were guided to take any seats that they wanted. It was observed that some groups of test-takers who were friends tended to group together.

Proctors

Most of the proctors for the three tests were instructors of English supplementary courses, or teaching assistants as well as professors at the English Department, and students at ISU who were legible for working on campus. All of them had no training on their duties, or about the test, and generally followed what Yoo-Ree Chung – the main person in charge told.

Giving instructions

There was an instruction sheet printed for the instructor to read throughout the test. The instructor received the sheet and read it along the test administration. The sheet can be found in the EPT administration and result processing manual (2010). For the listening test, the proctor made sure the test-takers to turn to the Listening section in the test booklet at the designated time, and to transfer their answers on the answer sheet correctly. The listening test was speeded by a recording.

Scoring and reporting EPT Listening test results

Right after each test, Yoo-Ree Chung collected all the answer sheets and took them to the Solution Center. The answer sheets including both reading and listening sections were then scanned, and the EPT Reading and Listening test results were processed immediately and completed within a couple of hours. The results were sent to Yoo-Ree Chung via email including the raw data (i.e. individual students' responses to each question) and students' raw scores on the reading and listening sections. Yoo-Ree Chung then transferred the EPT Reading and Listening results to new files to make placement decisions and reporting the results to relevant parties

including academic advisors and the Graduate college. All the results were finally imported into the Test bank for recording, and test revision.

2. Methodology

This section is to present which methods are chosen to address the stated research questions. First, the rationales for the selection of instruments with detailed descriptions are provided. Then, some main sketches on how each instrument is used, and how the data are collected and analyzed, follow.

2.1. Methods

Adopting the argument-based validation approach to examine validity of the EPT Listening test at ISU (Bachman, 1990, 2004; Chapelle, 1999; Kane, 1992, 2002, 2004), I decided to use a mixed method to address the four stated research questions corresponding to four main inferences (Domain Description, Observation, Generalization, and Explanation) in the proposed structure of the validity argument for the Listening EPT test. In specific, the evidences from both qualitative and quantitative methods, which are to back certain assumptions in each of the four main inferences in the validity argument of the EPT Listening test at ISU, are combined and structured to judge the plausibility of the test score interpretation and use. This mixed method is agreed to provide a proper insight into the validity issue, and strengthen the argument made as each method has its own strengths and drawbacks (Bachman, 1990, 2004; Chapelle, 1999; Douglas, 2009; Messick, 1989; Kane, 2002, 2004).

Based on the review of major methods in collecting evidences for validity of a testing in the specific context of my study (Bachman, 1990, 2004; Brown, 1996; Chapelle, 1999; Douglas, 2003; 2009; Messick, 1989), I would like to employ some instruments. The qualitative method comprises of test analysis including test-task analysis, and test-item analysis while the quantitative method in use is statistical analysis consisting of descriptive statistical analysis, reliability report, and correlation analysis.

2.2. Description of the instruments for the study

2.2.1. Test analysis

Two specific instruments of test analysis in use are test task analysis, and test item analysis. While the test task analysis is purposed to provide a qualitatively analytical insight into the content and characteristics of the EPT Listening test of Fall 2010 (set C2), the test item

analysis aims to give some quantitative evidences on the quality of test items in the EPT Listening test of Fall 2010 (set C2).

Test task analysis

The test task analysis basically employs the framework of listening task characteristics proposed by Buck (2001) which is claimed to base on the one given by Bachman and Palmer (1996). According to Buck (2001, p. 106), this framework is intended to function as a checklist for comparing test tasks with target-language use tasks which cover main aspects of a language test task. This comparison is also considered as a means of investigating task authenticity, as well as an aid to the development of new tasks. However, due to the unavailability of information on the design and development history of the Listening EPT test in Fall 2010 (Set C2), some main categories in the framework will be under examination in my investigation. The brief framework used to analyze the EPT Listening test booklet (Set C2) is presented in Table 7 below. More details on how to analyze the characteristics of the EPT Listening test of Fall 2010 in my study are provided in Appendix 2.

Table 8: The brief framework for analyzing the EPT Listening test at ISU in Fall 2010 (set C2) (adapted from Buck, 2001, p. 107)

Characteristics of the setting: It consists of all the physical circumstances under which the listening takes place. The physical conditions include all the material and equipment resources needed for a listening test. Participants need to be provided with proper instructions and the best conditions in order to have the best performance.
Characteristics of the test rubric: The test rubric includes those characteristics of the test that provide structure to the test and the tasks such as instructions, test structure, time allotment, scoring method.
Characteristics of the input: The input into a listening task consists of listening texts, instructions, questions, and any materials required by the task. Some aspects are under examination: (1) format, (2) language of input, (3) topical knowledge.
Characteristics of the expected response: There are two main aspects of an expected response of interest: format of expected response and language of expected response.
Relationship between the input and response: There are a number of aspects to look into the relationship between the input and response including reactivity, scope, and directness of relationship between the input and response.
Question types/formats:
Question types: There are two main question types used in a listening test: (1) Comprehension questions, and (2) Inference questions.
Test question format: Three common question formats include (1) short-answer questions, (2) multiple-choice questions, and (3) true/false questions.

Test item analysis

Item analysis is described as **"**the systematic evaluation of the effectiveness of the individual items on a test" (Brown, 1996, p. 50). It is purposed to select the best items that will remain on a revised and improved version of the test, or to investigate how well the items on a test are working with a particular group of students. Item analysis can take numerous forms, but when testing for norm-referenced purposes, four types of analyses are typically applied: item format analysis, item facility analysis, item discrimination analysis, and distraction efficiency analysis (Brown, 1996, p. 50). While the first one is qualitative, the other three are quantitative.

Some guidelines for carrying out each kind of item analysis are provided which is of great importance to my study. First, item format analysis focuses on the degree to which each item is properly written so that it measures all and only the desired content. Such analyses often involve making judgments about the adequacy of item formats. Second, item facility analysis employs item facility (IF) which is a statistical index used to examine the percentage of students who correctly answer a given item. Next, item discrimination analysis involves the production of item discrimination (ID) which indicates the degree to which an item separates the students who performed well from those who performed poorly. The last one – distractor efficiency analysis is to produce distraction indices indicating how well a certain test item distracts test-takers from getting the correct answer.

For a more informed quantitative analysis of test items in the EPT Listening test at ISU in Fall 2010, I also decided to refer to the proposal of some critical values for evaluating test item facility and discrimination given by Siriluck Usaha (1996) in the investigation into the reliability of the Suranaree University English Placement Test. These values are nearly similar to those suggested by Ebel (1979, p. 267), except for the last group of poor items which include all the items whose item discrimination indices are lower than or equal to 0.19.

Table 9: Criteria for item selection and interpretation of item difficulty index

Type	**Index of Difficulty**	**Evaluation of Difficulty**
1	0.80-1.00	Too easy
2	0.60-0.79	Rather easy
3	0.40-0.59	Moderately difficult
4	0.20-0.39	Rather difficult
5	0.00-0.19	Too difficult

Table 10: Criteria for item selection and interpretation of item discrimination index

Type	Index of Discrimination	Evaluation of Discrimination
1	0.60-1.00	Very good items
2	0.40-0.59	Good items
3	0.20-0.39	Reasonably good but possibly subject to improvement
4	0.10-0.19	Marginal items, usually need and subject to improvement
5	0.00-0.09	Poor items, to be rejected or rewritten

2.2.2. Statistical analysis

Most of testing experts agree that testing much involves scores or numerical data (Bachman, 1990, 2004; Brown, 1996). Thus, statistical analyses of test scores provide a lot of information about a test such as reliability, and other empirical evidences on validity. Three of the most common instruments of statistical analyses for norm-referenced tests have been introduced to language testers, teachers and administrators in a number of books (Bachman, 1990, 2004; Brown, 1996; Douglas, 2009). They are: (1) reporting descriptive statistics (describing and interpreting test results), (2) producing test reliability, and (3) doing correlation analyses.

(1) Reporting descriptive statistics

Descriptive statistics are described as numerical representations of how a group of students performed on a test (Brown, 1996, p. 102-109). In other words, such statistical analyses help to visualize test-takers' performances on the test in support to the understanding of complex patterns in test behaviors of test-takers.

(2) Producing test reliability

In general, the test reliability is described as the extent to which the results can be considered consistent or stable (Bachman, 1990, 2004; Brown, 1996; Douglas, 2009). Reliability is defined as a basic requirement of a valid test, and refers to the consistency of measurement. Numerous strategies with statistical tools have been introduced in order to investigate the issue of consistency in measurement. For norm-referenced tests, testers use reliability coefficients and the standard error of measurement (SEM**)** to examine the reliability of a test.

A reliability coefficient can be interpreted as the percent of systematic, consistent, or reliable variance in the scores on a test. Its value ranges from 0 to 1 (Brown, 1996, p. 193). There are three basic strategies to estimate the reliability of most tests: the test-retest, equivalent-forms, and internal consistency strategies. Of the three estimates, internal-consistency estimates are the

ones which are mostly-used by language testers because this type of reliability has the advantages of being estimable from one administration of a single form of a test. There are some different ways to estimate internal-consistency reliability: split-half reliability, Cronbach alpha, Kuder-Richardson formulas (KR-20, and KR-21).

Brown states that the KR-20 strategy is the single most accurate of these estimates. However, the other three approaches have advantages that sometimes outweigh the need for accuracy. For instance, the split-half version is more meaningful in explaining how internal-consistency reliability of a test works. The KR-21 formula has the advantage of being quick and easy to calculate. Cronbach alpha should be chosen to apply to tests with weighted items whereas the KR-20 can only be applied when the items are scored correct/incorrect with no weighting scheme of any kind. Again, when accuracy is the main concern, the KR-20 formula is highly recommended. Thus, in my study, I decided to use the KR-20 formula which is calculated based on the number of items, the mean, and the standard deviation on a test.

(3) Correlation analyses

Considered to be one of the most valuable sets of analytical techniques, the purpose of correlation analyses in language testing is to examine how the scores on two tests disperse, spread out the students in order to know whether that relationship is statistically significant as well as logically meaningful (Brown, 1996, p. 151). The concept of correlation coefficient with a number of its types has been examined (Bachman, 1990, 2004; Brown, 1996; Douglas, 2009). A correlation coefficient is defined as a statistical estimate of the degree to which two sets of scores vary together. Its value ranges from -1 to 1, and approaches 0 when there is absolutely no relationship between two sets of scores or numbers.

Some common types of correlation coefficient are Pearson product-moment correlation coefficient, Spearman rank-order correlation coefficient, and point-biserial correlation coefficient each of which has its own restrictions on usage. Accordingly, the Pearson product-moment correlation coefficient is chosen to compare two sets of interval or ratio scale data (Brown, 1996, p. 156). On the other hand, the Spearman rank-order correlation coefficient is used when two sets of scores are ordinal or nominal scales (Brown, 1996, p. 172). Finally, the point-biserial correlation coefficient is applied when examining the relationship between a nominal and an interval scale (Brown, 1996, p. 167).

For my study, the selection of which correlation coefficient will be decided on the descriptive statistical analysis of the score data on the EPT Listening test in Fall 2010, and that of the score data on other standardized English tests.

2.3. Procedures for data collection and data analysis

Data collection

Due to the scope of the three-month long study, and the unavailability of the information and data of the EPT test history at ISU, the study focuses on the EPT Listening test in Fall 2010 in order to have an insightful view into some main different aspects of the test construing the validity argument of the EPT Listening test. All the relevant and accessible data of the EPT test at ISU in Fall 2010 and its test-takers' scores on the internationally-standardized tests (TOEFL iBT and TOEFL pBT) for their admissions to the ISU are collected for analysis.

In specific, three main sources of data were used. First, all the EPT test manual and other documents of the EPT test in Fall 2010 including the Listening test specification, the test booklet (Set C2) and the EPT test result summary report were retrieved. The second source of data collection were the EPT Listening score set of all the test-takers in Fall 2010, and its placement results. Finally, the study collected the score sets of 395 EPT test-takers at ISU in Fall 2010 on the internationally English language standardized test developed by the Educational Testing Service (ETS) including both TOEFL iBT, and TOEFL pBT. All these sources of data were provided by the EPT test coordinator – Professor Volker Hegelheimer, and the EPT test assistant –Yoo-Ree Chung. They also provided some information on how the test booklet (set C2) was designed and developed.

Data analysis

Based on the detailed description of the instruments used in my study above, the three sources of data were processed using three main instruments as following. First, the test analysis employed the first source of data for analysis. In specific, the test booklet of Set C2 was examined on the basis of Buck's framework of listening test task characteristics (see Table 7). An inter-reliability index by two evaluators was run to seek a backing evidence on the test analysis results retrieved. The second evaluator was asked to analyze 25% of the total number of test items in the test booklet. In order to ensure the objectivity in selecting a variety of test items, the last lecture was specifically chosen for the analysis by the second evaluator.

The test analysis results of Set C2 would also be triangled with those produced by examining other data sources including the EPT Listening specification, the EPT test manual (May, 2010), as well as other relevant theoretical foundations for testing academic listening in second language.

Next, the score set of 556 test-takers of the EPT Listening Fall 2010 administration was sorted out in order to run some descriptive statistics, and to produce some reliability estimates.

Finally, the correlation analyses were intended to yield some inferential statistics about the interrelationship in measurement between the EPT Listening test with another internationally-standardized language test (TOEFL). Because there were two versions of the TOEFL test offered by the ETS (TOEFL paper-based test and TOEFL Internet-based test) with their different availability of Listening component scores, numerous correlation analyses were carried out: (1) between the EPT Listening Fall 2010 score set and the TOEFL iBT Listening score set, (2) between the EPT Listening Fall 2010 score set and the TOEFL iBT total score set, (3) between the EPT Listening Fall 2010 score set and the TOEFL iBT total score set including the converted TOEFL pBT scores, and (4) between the EPT Listening Fall 2010 score set and the TOEFL pBT score set.

Some steps to carry out the correlation analyses were taken. First, after doing an insightful and critical review of the selected tests and their score sets, some theoretically-based hypotheses on the correlations between them were made. Based on the examination on the nature of any two chosen test score sets, an appropriate correlation coefficient were adopted. Afterwards, the correlation coefficients were interpreted in terms of statistical significance and meaningfulness. All the statistical analyses were done with the assistance of Microsoft Excel 2007, and JMP 8 software.

In general, by analyzing statistically the given test data and interpreting the collected results, I purpose to examine how the scores of three different tests are related and figure out what factors influence the relationships among them. Then, an insight into the validity and the reliability of the EPT Listening test in Fall 2010 (Set C2) at ISU will be given as being compared with the widely acclaimed test – TOEFL, and vice versa, which is of great importance to the test developers in particular and test-takers or test-users of the EPT test at ISU. In other words, the correlation analyses can help to answer the question whether the three tests measure the same thing or not.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter comprises of two main parts. The first part summarizes the main results of three main analyses used in the study including the EPT Listening test analysis (Set C2), the EPT Listening Fall 2010 test score analysis, and the correlation analysis between the EPT Listening scores and the TOEFL Listening scores of the EPT test-takers in Fall 2010. Next, the second part presents the main findings withdrawn from the discussion of the results in the first part. The expected outcome of the study is the justification of the validity argument for the EPT Listening test of Fall 2010, which helps to suggest future revisions as well as relevant evidences to strengthen the validity argument for the EPT Listening test in particular, and the EPT test in general.

1. Results of the study

1.1. Analysis of the EPT Listening test of Fall 2010 at ISU (Set C2)

The EPT Listening test analysis consists of two main results, which are produced by a test task characteristic analysis, and a test item analysis.

Test task characteristics analysis

Relevant data sources for the test task characteristics analysis include the EPT Listening test specification, the EPT Listening test booklet (Set C2) with its accompanied recording, and other reference sources such as the EPT test manual, and the developers of the test booklet. The analysis results of the EPT Listening test in Fall 2010 at ISU are reported under each category given in Buck's framework of listening test task characteristics (2001). The framework covers (1) characteristics of the setting, (2) characteristic of the test rubric, (3) characteristic of the input, (4) characteristic of the expected response, (5) relationship between the input and response, and (6) question types and format. However, due to the unavailability of information on the development of the EPT Listening test booklet in Fall 2010 (Set C2), not all the categories are carefully examined. In fact, two categories (5 and 6) are chosen to be the focus of the analysis. Noticeably, the inter-reliability index by the two evaluators on the eight out of thirty test items in Set C2, specifically in lecture 4 reaches 0.7, which helps to confirm the analysis results acceptable and reliable.

(1) Characteristics of the setting

Based on the report of the EPT test in Fall 2010 in Chapter 3, some observations on the characteristics of the setting for the EPT Listening test in Fall 2010 can be made here. First, in terms of physical characteristics, all the material and equipment resources for the listening test were assured to provide a good condition for the test-takers' optimal performance. For instance, the quality of recordings was checked by the test developers, coordinators and the test instructors before the test. So were the players and loudspeakers in the room, which ensured every test-taker could hear the same regardless of his or her seat. Next, the test-takers were provided with proper instructions and supports if needed in order to have the best performance. For example, the administrators used a projector to demonstrate what were supposed to do in order to help the test takers to follow the instructions correctly. Also, most of the students had been informed of the test before arriving at the ISU and could learn about it through information available online. Finally, a number of different times and dates for the EPT test in Fall 2010 were offered for the test-takers to choose. However, the students were supposed to take the test right after their arrivals, which might have caused some disadvantages to those who suffered from jetlags or arrived late.

(2) Characteristics of the test rubric

The examination of some main characteristics of the test rubric on the EPT Listening test specification and the test booklet (Set C2) yielded some evidence in terms of test structure, test instructions, time allotment, and scoring method as follows.

First, a close look at the test specification of the EPT Listening test gives some information on the test structure and on the development as well as design of the EPT Listening test at ISU in general, the EPT Listening test booklet used in Fall 2010 (Set C2) in specific. The test specification for the EPT Listening test at ISU was created on 22nd March 2007, and is claimed to be based on the framework of academic listening given by Buck (2001), and the hybrid of the test-task characteristics (Bachman, & Palmer, 1996) and Davidson and Lyn's model (See Appendix 1).

Despite being noted as a draft, the specification covers some main contents needed for developing and designing the test. It is structured into three main parts. The first part presents the skills to be measured by the test. As being stated, the EPT Listening test is intended to measure academic listening comprehension and conversational listening comprehension. Five specific

sub-skills under examination in the test are also specified, including synthesis of information in the text, recognition and recovery of information in the form of specific details, recognition of opinions, recognition of inferences drawn from statements and information presented in the text, and identification of the meaning of key vocabulary items in the text.

The second part gives some guidelines on the content and the format of test tasks, and test questions in the test itself. In specific, it describes the format of the input for the listening test, prompt attributes, item type descriptions, and response attributes. For instance, the length of the input for a lecture should be 600 words long, and other channels such as video, images or charts are in use. For designing questions, it is suggested that four options need to be provided whereby one choice represents the correct answer, one choice is plausible (incorrect given the context), one choice is too narrow, and one choice is too broad. The correct answer needs to be marked with an asterisk (*). These questions fall into one of the three groups: (1) basic understanding, (2) pragmatic understanding, and (3) connecting information. Moreover, the distribution of the number of questions among these types for each listening text is given. Accordingly, an EPT Listening test is suggested to consist of 10-12 multiple choice questions with four choices including 5-6 basic understanding questions, 2-3 pragmatic understanding questions, and 2-3 connecting information questions. Noticeably, the specification specifies the allowance of note-taking for the test-takers during the test.

The third part is claimed to include attachments of sample items. Nevertheless, none is found which makes the test specification incomplete.

More insightful observations about test structure, and other aspects of the test rubric including test instructions, and scoring method are gathered as carrying out an investigation into the real test booklet (Set C2) used in the EPT Listening Fall 2010 administration.

The EPT Listening test of Fall 2010 was administered by a recording which includes instructions, four listening texts followed by a series of questions, and response pauses. The examination of the recording helped to reveal the time allotment of the test, which is determined by the sequence of texts and tasks. The test lasts for 50 minutes in total, 20 minutes of which is response time. On the average, each answer takes about 40 seconds to answer. In addition, the time allotment of response time among the four lectures is found to correspond to the length of their listening texts. The specific allotment of response time among the four lectures in the EPT Listening test (set C2) can be summarized here: Lecture 1 (4 minutes – 537 words in 2:58

minutes), lecture 2 (4 minutes – 358 words in 2:33 minutes), lecture 3 (5 minutes – 478 words in 3:28 minutes), and lecture 4 (7 minutes – 1299 words in 7 minutes).

Despite being administered by a recording, the test administration was reported to proceed smoothly. The instructions were recorded by an American female speaker, and were found to be clear, and simple. The instructions comprise of a number of information, including an introduction about the test purpose, its main components, time allowance, and some brief guides on how to do the test. Noticeably, the instructions are given in both spoken and written forms, which are printed in the test booklet and are presented in Table 10 below. The comparison of these two instructions suggests that the two instructions are complementary while the spoken instructions include more details than the written ones. For example, the spoken instructions contain a notice on how to take notes to do the test, which is not included in the written instructions; specifically '*you don't need to take notes all the lecturer says, but main ideas and concepts'*. Another observation is that three critical pieces of information in the written instructions are formatted to be bold and italicized for visual effects, which is found to be very helpful.

Table 11: EPT Listening test instructions (Set C2)

Listening test *(Spoken Instructions)* This listening test will indicate how well you understand spoken English in some typical situation that you may encounter in university. The listening test comprises of four parts. For each part, you'll watch a video lecture, or an interview talk. While watching it, you'll take notes on a separate sheet. You may answer the questions using your notes. Record your answer on the computer form beginning with the item 51. Do not mark on the test booklet. This test will take approximately 50 minutes. Now put your computer form aside, and take the note-taking sheet for the first lecture. You'll hear a lecture about Team Composition. You don't need to take notes all the lecturer says, but main ideas and concepts. You'll have only one chance to listen each. Ending: this is the end of the lecture. Answer the questions on part D on the test booklet and record your answers on the computer sheet you might use your notes to answer them. Do not write on your test booklet. *(Written instructions)* This listening test will indicate how well you understand spoken English in typical situations that you may encounter at the university. The listening test consists of **four lectures**. For each lecture, you will take notes on a separate note-taking sheet that you will be given and then answer questions using your notes. You will record your answers on the computer forms, <u>starting with item 51.</u> This test will last approximately **60 minutes.** Now, put your computer form aside and look at the note-taking sheet to take notes for the first lecture.

In terms of scoring method, the EPT Listening test (Set C2) uses the multiple-choice format for efficiency and convenience. Accordingly, each test-taker is distributed with a computer answer recording form, and is instructed to transfer his/her answer onto the computer answer recording sheet. The computer answer forms will be automatically scanned and scored by an authorized technician. The listening score is based on the number of correct answers without differences in score weighting among them. No information on how to score each question is found in the test instructions, which is, however, expected to not affect the final listening score of the EPT test-takers.

On the other hand, some concerns arose as looking into the EPT Listening test of Fall 2010. First, its test specification still lacks significant information on the organization of the test, or its general structure. Moreover, as comparing the specification with the real test booklet used in Fall 2010 (Set C2), some mismatches can be seen between them. For example, while the EPT Listening section is specified to contain both academic lectures and short conversations, the real test comprises of four academic lectures without any conversations. Furthermore, the EPT Listening test set (Set C2) is inspected to have some shortcomings in its instructions. The test does not offer any example to prepare the test-takers before they do the test besides a notice on which cell to start the listening section on the computer answer recording form. With the assumption that all the test takers are familiar with the multiple-choice format, the test writers fail to provide the test-takers with explicit criteria for choosing an answer to a multiple-choice question.

(3) Characteristics of the input

The analysis of the input for the EPT Listening Fall 2010 administration lead to some brief descriptions about its format, topical knowledge, and language of input which are presented in Table 11 below.

The listening texts in the EPT Fall 2010 test are all authentic videos taken from reliable resources on the Internet. All the four lectures in the EPT Listening test (set C2) were given by native English speaking professors in either the U.S or Britain, which are thus expected to be highly representative of the target spoken language in the U.S colleges, and universities. To be specific, lectures 1 and 3 are delivered by two professors at Standford University, lecture 2 by a professor at the ISU, and lecture 4 by a professor from University College in London.

There are some interesting remarks about the format of the input into the EPT Listening test in Fall 2010 (Set C2). All the four listening texts in the test have a lead-in by a narrator introducing the main topic of each lecturer and preparing the test-takers to listen with the integration of other channels in order to measure the students' academic listening performance effectively. For example, the first three lectures all start with a slide presenting the presenter's name, the topic of the lecture followed by a video employing captions, and images along the lecture. In fact, the captions appear to be quite small on the screen for the back-rows in the testing room, but they are not intended to be read by the test-takers. Except the fourth lecture, which was added to Set C1 to make Set C2, all the three lecturers in Set C1 meet the length requirement in the specification of the EPT test, ranging from 358 words to 537 words.

Table 12: Some descriptions about the four listening texts in the EPT Listening test in Fall 2010 (Set C2, n=30)

Lecture No	Length (words)	Duration (minutes)	Channel	Topic
1	537	2:58	Audio, video, caption	Team composition
2	358	2:33	Audio, video, caption	Research in plant pathology
3	478	3:28	Audio, video, caption	Car driving simulation
4	1299	7	Audio, video, caption	How the internet enables intimacy

As can be seen in Table 11, a wide range of topical knowledge is covered in the EPT Listening test booklet. Each of them taps on a different academic field, specifically social science (lecture 1, lecture 4), natural science, technology and engineering (lecture 2, lecture 3). As the lecturers are available on the Internet for educational purposes without any restrictions on the viewers, the contents of these listening texts are expected to be not too technical.

With the scope of the study, some general linguistic and audio characteristics of the input into the EPT Listening test in Fall 2010 (Set C2) can be described. In terms of linguistic features, while lectures 1, 2, and 4 are mono-logic, lecture 3 is more interactive as a news report. Various sentence types and grammatical structures are found in the listening texts. Next, in terms of audio features, all the speakers are native speakers of English with two major accents, i.e. Northern American English, and British English. A significance about the audio features of the

listening texts is the inclusion of the Iowan accent by a professor in a lecture about pathogen in the context of Iowa.

(4) Characteristics of the expected response

As the format of the EPT Listening test (Set C2) is multiple-choice in which the test takers are given four choices, the answers are partially structured with the dichotomous scoring rubric (Right/Wrong; 1-0). All the questions and options are in English. Therefore, it does not cost much effort of the test-takers to structure the answer, and those of test-administrators to score their answer.

(5) Relationship between the input and response

Some aspects of the relationship between the input and response under examination s are directness, and interactiveness which in this study refer to the dependency on the content of listening texts, and the employment of listening skills and relevant academic sub-skills to succeed on the test. The detailed analysis of all the test items in the test set (Set C2) can be found in Appendix 5. The summary results show that twenty two out of thirty test items in the test booklet were evaluated to have high passage dependency, and interactiveness. In other words, in order to be highly probable to make a correct-choice for these items, the test-takers have to rely on their comprehension of the listening texts instead of their merely background knowledge, as well as to be fluent with relevant academic listening skill such as note-taking to catch major or minor details, connecting ideas, and synthesizing information.

A detailed analysis of the test items in the EPT Listening test of Fall 2010 (Set C2) about their engagement of different academic listening sub-skills, strategies, or areas of language knowledge of the test-takers are also presented in Appendixes 4 and 5. Some examples can be given here to illustrate how direct, and interactive the test items in the EPT Fall 2010 Listening test booklet are. Question 63 in the second lecture is an example of passage-dependency, and interactiveness in the academic context. The question checks a detail in the first section of the listening text, which requires the test-takers to take good notes, and to understand the presented information in order to choose the best answer.

Question 63: The speaker and his associates developed the car simulator in order to create situations that would:

(A) Eliminate physical danger while giving a person practical experience on the road.

(B) Increase sociologists' understanding of how people behave in a car. It requires both a speaker and a listener.

(C) Assist auto manufacturers' future design of features a customer may want in a car.
(D) Allow a person who has never driven before the sensation of driving in a variety of conditions.

Next, in order to answer Question 58 in the first lecture, the test-takers have to comprehend the section of the listening text, synthesize information in order to make an inference about the speaker's emphasis on the value of seed-borne pathogen research, which is highly representative of academic skills for successful communication. Noticeably, all the sub-skills described in the EPT Listening test specification (i.e. synthesis of information, recognition and recovery of information in the forms of specific details, recognition of opinions, recognition of inferences drawn from statements and information presented in the text, and identification of the meaning of key vocabulary items in the text) are observed to be included in one of these test items.

Question 58: The scientist says "microtoxins are natural metabolized fungi".
(A) To summarize his speech
(B) To define a technical term
(C) To support an opinion
(D) To provide an example

However, the other eight items out of thirty test items in Set C2 (Q62, Q65, Q66, Q73, Q75, Q76, Q79, Q80) were assessed to either have lower passage dependency, or low representativeness of knowledge, skills and abilities. Five out of these seven items (Questions 62, 65, 66, 73, and 76) were found to have a high interactiveness but low passage dependency while the other two (Questions 79, 80) were seen to have a high passage dependency but low interactiveness. On the other hand, Question 75 was evaluated to have neither high interactiveness nor directness. For the first group, the first five items were evaluated to engage the test-takers' highly representative academic sub-skills such as synthesizing, or inferencing, but the test-takers can use their background knowledge or intelligence to have the correct answer. For example, Question 62 is a comprehension question about some agricultural products in Iowa, whose four given choices are quite easy and clear for those who have already learnt about Iowa. Thus, the test-takers might have a correct answer based on their background knowledge about Iowa. On the other hand, Question 66 illustrates how the test-takers can use their intelligence without listening comprehension to answer it correctly. In specific, three out of

the four given choices are relevant, but too specific while the correct option is found to restate the given statement the most closely and sufficiently.

Question 62: Which of the following is NOT true about the relationship between agriculture and the economy of Iowa?

(A) The economy of Iowa heavily relies on the productivity of farming.
(B) Iowa's main crops are soybeans and corns.
(C) The amount of seed production for soy beans is very small in Iowa.
(D) Seed production is very important for the success of Iowa farmers.

Question 66: The speaker claims there is a "tradeoff" between knowledge as helpful and knowledge as harmful. In saying this he is:

(A) Highlighting the risks involved in using car simulation vs. advantages of real-life road experience.
(B) Warning consumers of the hazards of having GPS in their automobiles.
(C) Urging the listener to get involved in research on how to improve current technology in cars.
(D) Raising the issue of benefits vs. drawbacks of having knowledgeable cars that track our personal information.

In contrast, the other two items (Question 79, 80) are found to be highly dependent on the details given in the listening text (Lecture 4). Nevertheless, they both involve two specific details in the lecture which are not much essential to the main ideas, and the topic of the lecture. Hence, answering these two items correctly may not show how well the test-takers can typically perform in another similar academic setting for effective communication.

Question 79: The speaker probably thinks that the reported percentage of the people who do personal email at work is conservative based on:

(A) Her own research results with mobile phones.
(B) The report of an anthropologist's Facebook study.
(C) The results of the research conducted by the U.S Army.
(D) Her interviews with several close couples.

Question 80: According to the talk, the isolation of the private sphere from the professional domain began approximately----------years ago.

(A) 15
(B) 50
(C) 115
(D) 150

Finally, all the questions and answers in the EPT Listening test (Set C2) are provided in the written format only, which requires the test-takers' ability to read fluently in order to have a successful performance on it. This feature which bears a weak relation to the listening construct

might create influence on the validity of the EPT Listening test score interpretation. The disadvantage of the provision of the 'only written' formatted listening questions can be better seen through the following two examples (Questions 74 & 75) in the test booklet (Set C2). So as designing the test, the test-takers are assumed to be able to read English in order to understand the given questions, the given choices, and to choose the best answer. Significantly, for Question 75, the students can choose the correct answer based on their reading comprehension without the comprehension of the listening text.

Q.74. What does the speaker mean by "rituals" in this talk?
(A) Religious procedures
(B) Prescribed orders for a *ceremonie*s (misspelled in the test booklet)
(C) Habitual daily routines
(D) A series of actions

Q.75. What does the speaker mean when she says that that children are educated to "do (this) cleavage" between professional lives and personal lives?
She means that they are taught to……
(A) distinguish professional lives from personal lives
(B) connect professional lives and personal lives
(C) replace professional lives with personal lives
(D) prefer professional lives to personal lives

Another problem is with the two test items in the last listening lecture (Lecture 4). In specific, the correct answers for Questions 76 and 80 both rely on the same piece of information, which relates to how long the isolation between the public and private spheres has been.

(6) Question types/formats

The EPT Listening test in Fall 2010 (Set C2) adopting the multiple-choice format comprises of two main types of questions: comprehension questions and inference questions (Buck, 2008, Chapter 5). Based on the question classification framework by Shohamy and Inbar (1991), the results of the scrutiny of the test items in the test booklet are summarized in Table 12 below.

Table 13: Summary of analysis results about question types for the EPT Listening test of Fall 2010 (Set C2, n=30)

Lecture	**Total**	**Question types**							
		Comprehension			**Inference**				
		Global	Local	Trivial	Main idea	Pragmatic/ sociolinguistic implication	Pragmatic/ Sociolinguistic purpose	A gist/or unclearly stated section	Inference of word-meaning
1	**6**	Q52, Q53, Q56			Q51	Q54, Q55			
		3			***1***	***2***			
2	**6**	Q57, Q62	Q60			Q58, Q61		Q59	
		2	***1***			***2***		***1***	
3	**8**	Q70	Q63, Q68				Q64, Q66, Q69	Q67	Q65
		1	***2***				***3***	***1***	***1***
4	**10**	Q72, Q73, Q76, Q78		Q79, Q80	Q71	Q75		Q77	Q74
		4		***2***	***1***	***1***		***1***	***1***
Total	**30**	**10**	**3**	**2**	**2**	**5**	**3**	**3**	**2**

As can be seen from Table 12, the total thirty listening test items in the EPT Listening test booklet (Set C2) equally fall into the two groups. For the comprehension question group, ten out of fifteen questions in the four lectures are classified as global questions involving the test-takers' ability to synthesize information or draw conclusions. While three of the rest five questions fall into the local group asking the test-takers to locate details or understand individual words, the other two relies on trivial details in the listening texts. For the inference question group, the fifteen test items distribute well among different subgroups including asking for the main idea (2), asking for a gist of the spoken text, or a section of the text or an unclearly stated, but deliberately implied idea by the speaker (3), asking about pragmatic or sociolinguistic implication and purpose of the speaker (8), and asking about the word meaning in a specific context (2). Interestingly, the proportional distribution of these questions in the EPT Listening test in Fall 2010 (Set C2) is found to be the same as that given in the EPT Listening test specification. In addition, the fourth lecture is found to have the highest number of questions making up one third of the total number of questions in the whole EPT Listening test (Set C2).

Test item analysis

Two main test item indices (item difficulty, item discrimination) were used in the test item analysis for the EPT Listening test in Fall 201 (Set C2), the results of which were then examined based on the categorization scheme containing different ranges of item difficulty and discrimination indexes (Usaha, 1996). The specific test item indices of the thirty items in the EPT Listening test can be found in Appendix 3. Table 13 below presents the summary of the analysis results.

Table 14: Summary of item analysis results for the EPT Listening test in Fall 2010 (Set C2, n=30)

Difficulty	**Number**	**%**	**Discrimination**	**Number**	**%**
Too easy	1	3%	Very good items	6	20%
Rather easy	11	37%	Good items	7	23%
Moderately difficult	12	40%	Reasonably good but possibly subject to improvement	6	20%
Rather difficult	4	13%	Marginal items, usually need and subject to improvement	2	7%
Too difficult	2	7%	Poor items, to be rejected or rewritten	9	30%

In terms of difficulty, the Listening test (Set C2) had a fairly acceptable distribution among the five designated levels. Accordingly, the 'moderately difficult group' owned the largest number of test items (40%) while only 3 out of 30 test items (10%) were either 'too easy' or 'too difficult'. The other half of the total number of test items (50%) were evaluated to be 'rather easy' or 'rather difficult'.

Interestingly, in terms of discrimination, the EPT listening test (Set C2) contained more items of good discrimination (43%), a fair amount of test items needing improvement (27%). However, it included a high number of test items (30%) whose discrimination indices were very low. Those were evaluated to be 'poor', and needed to be rewritten or replaced.

A further investigation into item distraction efficiency of the test items with low discrimination gave another insight into the quality of the EPT Listening test of Fall 2010 (Set C2). To be specific, all the items with discrimination indices below to 0.25 were selected for item distraction analysis which was aimed to find out the frequency of selection by test-takers for

each choice given, and the selection frequency of the correct answer. Based on the results above, there were four listening items with discrimination indices under 0.25 in Set C2, which were chosen to be under examination. While two of the four listening items (Q66, Q77) had a quite acceptable difficulty level, the other two (Q76, 79) were shown to be too difficult. Three of them were defined to check listening comprehension while the other to test inferences. The distraction efficiency indices of these four items are presented in Table 14 below.

The examination of the results revealed some interesting facts. Two items out of these four items (Q66, Q77) had one of the four choices with very bad distraction (4% and 3%). In contrast, the other two (Q76, Q79) had a better distribution of selection rates among the four given choices, but their difficulty was quite high so that the percentages of the test-takers who got the right answer were unsatisfactory (less than 20%).

Table 15: Summary of item distraction analysis of four items with low discrimination indices (ID<0.25) in the EPT Listening test of Fall 2010 (Set C2)

Items	**Item Analysis**		**Distraction Analysis**			
	IF	**ID**	(a)	(b)	(c)	(d)
Q.66	0.678	0.229	44%	4%	10%	***42%***
Q.76	0.164	0.059	***16%***	21%	43%	19%
Q.77	0.507	0.241	34%	***51%***	12%	3%
Q.79	0.198	0.240	***20%***	25%	31%	25%

1.2. Statistical analyses of the EPT Listening test score of Fall 2010 at ISU

The basic statistical analysis of the EPT Listening Fall 2010 test administration covers a brief report of descriptive statistics of its test score set, and its reliability. They include: (1) some descriptive statistics of test scores (mean, mode, median, standard deviation, standard error measurement, and distribution), (2) reliability indices (KR-21, Cronbach alpha, split-half reliability), which are presented in Table 15 below.

Table 16: *Descriptive statistics of the test score set of the EPT Listening Fall 2010 administration (N=556)*

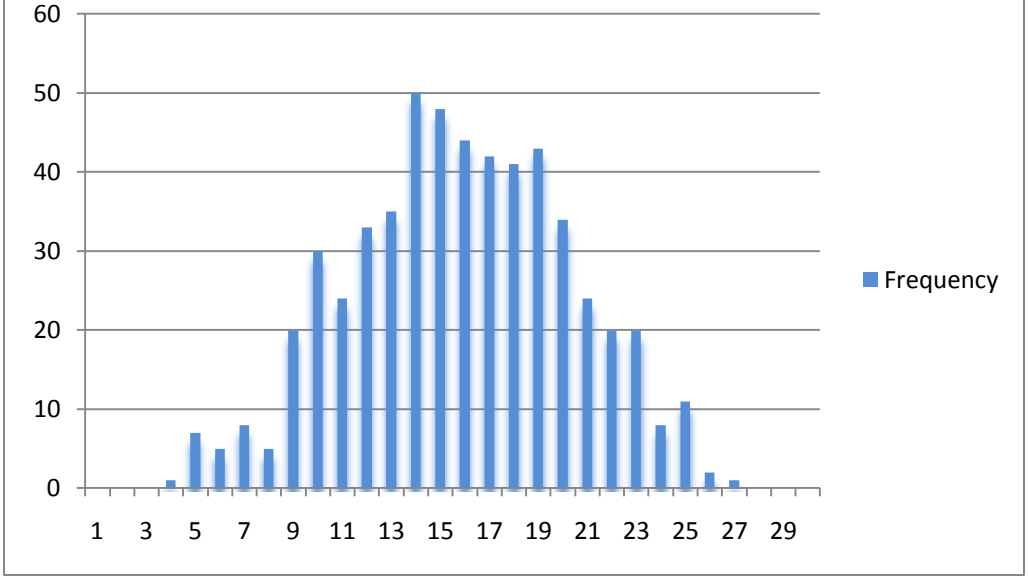
Listening (N=556, n=30)	
Mean	15.79
Mode	14
Median	16
Skewness	-0.08
Kurtosis	-0.42
Standard Deviation	4.52
SEM (KR21)	2.65
SEM (KR20)	2.48
R Split-half	0.99
Cronbach alpha (JMP)	0.69
KR20	0.69
KR21	0.66

As can be seen, the Fall 2010 Listening score set had fairly acceptable statistical results. While the highest possible score on the EPT Listening test is 30, the results of mean, mode and median in the Fall 2010 administration fell into the range of half of this score ranging from 14 to 16. Specifically, a number of the test-takers receiving a score of 14 and 15 in the EPT Listening Fall 2010 test were higher than those on other scores, which caused a slope on either side of the curve. However, the standard deviation was fairly big (about 4.5).

Based on these results, some characteristics of the test scores can be given, which are also visualized through the histogram below (see Figure 5). Accordingly, the EPT test score set of Fall 2010 had a fairly acceptable normality in distribution despite its not bell-curved shape. The histogram was negatively skewed as the distribution was seen to move slightly towards the right of the center line of the curve with the skewness value of approximately -0.079. Also, the score distribution was seen to be rather flat which was supported by its kurtosis value (-0,42) suggesting that the distribution did not have more extreme scores. In general, these skewness and

kurtosis values helped to show that the EPT Listening Fall 2010 score set had a reasonably normal distribution as they fell into the acceptable range from -2 to 2 (Bachman, 2004, p. 74).

Figure 5: Distribution of the score set of the EPT Listening Fall 2010 administration (N=556, n=30)



In addition, the EPT Listening Fall 2010 score set was found to be clearly separated. This observation was quantitatively supported by the percentages of listening test scores among different score ranges which were separated by one standard deviation. According to Douglas (2009, Chapter 5), the ideal bell-curve of a normal distribution should have a predictable ratio at various points among the one-standard unit scale between the minimum and the maximum values, which is 2.1%:13.6%:34.1%. The quantitative distribution result of the set score of the EPT Listening Fall 2010 administration was found to be quite satisfactory for a mid-stakes norm-referenced test. Accordingly, 34% and 37% of the test-takers got a score, which fell into the range within one standard deviation on either side of the mean score. Meanwhile, 11% and 13% of the listening scores ranged from one-standard deviation above or below the mean to two-standard deviations above or below the mean respectively. As being predictable, 2% and 3% of the rest of the test scores in the test set of the EPT Listening Fall 2010 test administration belonged to the range within three-standard deviations below or above the mean.

Finally, there are some statistical evidences on the reliability of the EPT Listening test in Fall 2010 (Set C2). As can be seen in Table 15, three different reliability methods were used to examine this aspect of the test booklet. Significantly, the internal consistency among test items

in the test booklet itself was quite strong with the split-half reliability of nearly 1. However, the KR20, KR21 and the Cronbach alpha indices showed another thing. These three indices were all lower than 0.7. In specific, the KR-20, which is claimed to be more reliable and precise than the others was around 0.6983. As suggested by testing experts (Douglas, 2009; Staynoff & Chapelle, 2005), the reliability is expected to reach 0.8 for a high-stakes test. Therefore, as being a mid-stakes test, the KR20 reliability of the EPT Listening Fall 2010 test was not too bad, but still called for the need to improve the test in order to attain the expected reliability of at least 0.7. Another reliability measurement in use was the standard error of measurement (SEM), which is also presented in Table 15. As can be seen, the SEM of the test score set of the EPT Listening Fall 2010 administration was quite substantial (about 2.65 (KR21); about 2.484 (KR20)) leading to the deduction that the a test –taker of the EPT Listening Fall 2010 administration (Set C2) would have a pretty wide probability of scores.

1.3. Correlation analyses of different score sets of the test-takers of the EPT Listening Fall 2010 administration at ISU

The correlation analyses in the study present the results of the investigation into four different correlations based on the test-takers' performances on the EPT Listening Fall 2010 test and on the TOEFL tests. The presentation follows the three-stage procedure of a correlation analysis. Accordingly, this part firstly introduces brief comparisons between the two versions of the TOEFL test (TOEFL pBT vs. TOEFL iBT), the TOEFL iBT test and the EPT test, and the listening section in the TOEFL iBT test versus the EPT Listening test (Set C2). Two major aspects under examination are test purpose and test formats. Based on the comparative review, some hypotheses about the correlation in the test-takers' performance on the two tests will be made. Next, after examining the characteristics of the available score data of the EPT Fall 2010 test-takers on the two tests, appropriate correlation coefficients will be decided. Finally, these collected statistical correlation evidences will be used to examine the given hypotheses in the final discussion.

(a) A review of the three tests under examination (TOEFL pBT, TOEFL iBT, and the EPT Listening Fall 2010 test at ISU

About the TOEFL

The Test Of English as a Foreign Language (TOEFL) administered by the ETS evaluates the proficiency and general understanding of the English Language for people whose first

language is not English. It is currently administered at test sites around the world in two different formats: the paper-based TOEFL (pBT) and the Internet-based TOEFL (iBT). In fact, the TOEFL test has evolved from paper based to computer based, and currently Internet-based. The first TOEFL iBT test was administered in the United States on September 24, 2005 and was launched in Internet Based test centers in Canada, Italy, France and Germany on October 22, 2005. It was rolled out to all other countries during 2006 in accordance with the schedule announced by ETS on October 1, 2005. In fact, the TOEFL pBT is offered for two purposes. One purpose of the TOEFL pBT is for placement and progress evaluation. Colleges or other institutions use the TOEFL pBT to test their students. The scores are not valid outside of the place where they are administered, but the college or institution accepts the TOEFL pBT that they administer as an official score. This pBT is also called an Institutional TOEFL. The other purpose of the pBT is to supplement the official TOEFL iBT in areas where Internet-based testing is not possible. The scores are usually valid outside of the place where they are administered. This pBT is also called a supplemental TOEFL.

As the TOEFL iBT scores mostly made up the TOEFL score set of the EPT Fall 2010, the focus of this review is to provide two comparisons including a comparative review between the two versions of the TOEFL test (TOEFL pBT vs. TOEFL iBT), and a comparison between the TOEFL iBT test including its listening test and the EPT Listening test, specifically Set C2. These comparisons are based on the model for test task characteristics and language ability by Bachman and Palmer (1996, p. 49 & p. 63) as well as the understanding of test purpose and test method by Staynoff and Chapelle (Chapter1, 2005). Both similarities and differences are found in terms of uses, task characteristics including setting, test rubrics, input, and expected responses, and constructs to be measured, which can be briefly described as below.

TOEFL pBT vs. TOEFL iBT

Based on the descriptions about the two tests published by the ETS on their official website http://www.ets.org/toefl/, a lot of changes in language measurement have been made in the new TOEFL format although the test users and its test uses are nearly unchanged. Accordingly, about a million people of all ages are reported to take the TOEFL test annually in order to demonstrate their English-language proficiency. They can be categorized into different groups of jobs, or purposes. For example, the TOEFL test-takers can either be students planning to study at a higher education institution, or students and workers applying for visas, ect. The test

is regarded to be a high-stakes test because its results will be used for a number of important purposes (ETS, 2011). First, it is stated that more than 7,500 universities and colleges in over 130 countries accept TOEFL scores as a language requirement for admission. Also, the TOEFL scores can be employed by other agencies and institutions at different levels of importance. For instance, the TOEFL score is one of criteria for the immigration departments to decide whether to issue residential and work visas for an applicant or not. At a lower scale of importance, it can be used by individuals to measure their progress in learning English.

On the other hand, some differences in what aspects language ability to be measured between the two versions of the TOEFL test can be clearly seen through the descriptions about them given by its test developer (ETS, 2011). In specific, the TOEFL pBT is a paper-based test that measures test-takers' ability to use and understand English in a classroom setting at the college or university level. It is also claimed to accurately measure how well a test taker can listen, read and write in English while performing academic tasks. However, despite the same goal of assessing test-takers' ability to use and understand English at the university level with the TOEFL pBT test, the TOEFL iBT test evaluates how well a test taker can combine his/her listening, reading, speaking and writing skills to perform academic tasks. As can be seen, the TOEFL iBT format focuses both on receptive skills and productive skills as including speaking skills, which is not tested in the TOEFL pBT version. Noticeably, the new TOEFL iBT test emphasizes on the integrated skills, which refers to the ability to integrate different language skills and sources of information for spoken or written production.

These changes in what to be measured of test-takers consequently lead to a lot of changes in test formats between the two TOEFL tests. A brief comparison in test structure and test format between these two tests can be summarized here. A detailed comparison can be found in Appendix 6. On the whole, the TOEFL pBT test has three parts: Listening comprehension, structure and written expression, and reading. In addition, the Test of Written English is an essay that is required to provide a writing score. The total TOEFL pBT score is based on a scale of 310-677. Differently, the Internet-based TOEFL (iBT) comprises of four parts: Listening, Speaking, Reading, and Writing. On the four-part TOEFL iBT, most of the questions are independent, but some of the questions are integrated. The total score is based on a scale of 0-120. Noticeably, while the TOEFL pBT total score does not include the Writing score, the Writing score on the TOEFL iBT is used to be computed to produce the total TOEFL iBT score.

Beside the introduction of the integrated tasks into the structure of the TOEFL iBT test, specifically the Speaking and Writing tests, some other significant changes can be observed. In terms of test format, the input in the Listening and Reading sections of the TOEFL iBT is more diverse with the inclusion of visual cues, and glossary. What is more, for question types and formats, a wider range of question types and formats is found in the TOEFL iBT for the Listening and Reading sections. For instance, apart from multiple-choice questions, the TOEFL iBT Listening and Reading tests comprise of matching questions in which the test-takers have to drag given choices to match each other. Next, the TOEFL iBT requires the test-takers to familiarize with computer and the Internet in order to follow the instructions and complete the tasks without difficulty while the TOEFL pBT is a traditional pen and paper-based test. Moreover, the time allowance for the two tests is significantly different. While four hours is needed to take a TOEFL iBT test, the TOEFL pBT takes about two hours.

TOEFL pBT Listening vs. TOEFL iBT Listening

Due to the scope of the study, this review will focus on similarities and differences in characteristics of the Listening section in the two tests which are summarized in Table 16 below. Beside some major common disparities mentioned earlier, with the aid of computer, the TOEFL iBT Listening section is seen to be more strictly delivered and speeded while comprising a wider range of question types than the TOEFL pBT Listening section. Also, the TOEFL iBT Listening version appears to emphasize on academic listening skills whereas the Listening section in the TOEFL pBT tends not to take this skill into consideration. Furthermore, with the integration of images, the Listening section of the TOEFL iBT test stimulates another dimension of metacognitive skills which is different from what the listeners use as listening to a recording without visual aids in the TOEFL pBT. However, to a large extent both these tests are claimed to measure academic listening of second language learners of English with similar sub-skills such as listening for pragmatic or social linguistic meanings, listening for main ideas, and inferencing.

Table 17: A brief comparison of the listening section in the two TOEFL tests (TOEFL pBT vs. TOEFL iBT)

Characteristic	**Items**	**Paper-based TOEFL (2 hours)**	**Internet-based TOEFL (4 hours)**
Test rubric	Test structure	50 questions	33-34 questions
		Three types of questions are presented in three separate parts. Part A has short conversations; Part B has long conversations and class discussion; Part C has mini-talks and lectures.	Two types of questions are presented in six sets: The first sets each have a long conversation. The next sets each have one lecture.
	Time allotment	Everyone taking the TOEFL proceeds at the same pace	The test-taker may control the pace by choosing when to begin the next conversation or lecture.
		The section is timed. At the end of the tape, the test-takers must have completed the section.	The section is timed. A clock on the screen shows the time remaining for the test-takers to complete the section.
	Instructions	In spoken format only	In both spoken and written formats
		Note-taking might not be allowed.	Note-taking is allowed while listening to the conversations and lectures
Input	Length	The talks and lectures are about 2 minutes long.	The lectures are about 5 minutes long.
	Format	There are no pictures or visual cues.	Each conversation and lecture begins with a picture to provide orientation. There are several pictures and visual cues with lectures.
Expected response		The test-takers can return to previous questions, erase, and change answers on their answer sheet.	The test-takers cannot return to previous questions. They can change their answer before clicking on OK.
Question types/formats		All of the questions are multiple-choice.	Most of the questions are multiple-choice, but some of the questions have special directions.
		Every question has only one answer.	Some of the questions have two or more answers.

TOEFL iBT vs. EPT test at ISU

This section presents two main results of the comparison and contrast between the TOEFL iBT and the EPT test. It first briefly reports some observed differences and similarities between the TOEFL iBT test and the EPT test in general by using the model for test task characteristics and language ability from Bachman and Palmer (1996, p. 49 & p. 63) as well as the first chapter of Staynoff and Chapelle on the understanding of test purpose and test method

(2005). A more specific comparative analysis of the Listening section between the two tests will be followed.

The scopes of inferences drawn from these two language tests and their types of educational uses greatly vary on the continua of inferences and uses given by Stoynoff and Chapelle (2005). According to them, the purposes of tests can be described in terms of "three dimensions that capture the important functions of the test" which include inferences made from the test, the uses of the test and the scope of the impact of the tests (Staynoff & Chapelle, 2005, p. 10).

The first dimension, concerning the inferences drawn from the test scores, is described on a continuum that ranges from specific (where connections are made to what is explicitly taught) to general (where the test measures general purpose language ability). Inferences about the test takers' language ability for both the TOEFL iBT and the ISU EPT test fall more towards the general side of the continuum, showing the test taker's academic and general English language ability. In specific, both the tests are to measure test-takers' language abilities in the English-medium academic context; however, the ISU Placement test is more specific in the academic environment in the U.S, specifically at the ISU.

The second dimension, which includes the educational uses or decisions made on the basis of test results, falls on a continuum that ranges from low to high stakes. The TOEFL iBT lies toward the high stakes side where admissions to university are involved (for this particular group of students) and the ISU EPT test falls towards the middle of the continuum as students are placed in ESL courses (99L, 99R, 101) based on the results they obtained in the EPT test.

For the third dimension concerning the scope of impact of the language tests, the TOEFL iBT has a broad impact which might influence the students, teachers, classes, programs, the institutions and society while the ISU EPT Listening test might have less of an impact on its users such as the society, institutions or the students.

Based on Bachman and Palmer's (1996) task characteristics framework, quick comparisons between the TOEFL iBT test and the ISU EPT test yielded both similarities and differences in terms of setting, test rubrics, input, expected responses, and constructs to be measured as following.

Some differences in setting between the two tests can be summarized here. The TOEFL iBT test was in most cases taken in the students' home countries while the EPT test was operated

at ISU. However, both the entire TOEFL iBT and the EPT test should be completed in one sitting. Second, the TOEFL iBT is administered to all the second language learners whereas the EPT test at ISU is taken only by the international students admitted to the ISU. Hence, the participants of the ISU EPT test have a more limited range of language ability and characteristics than those of the TOEFL iBT test. Also, whereas the students had unlimited opportunities to take the TOEFL iBT test to reach their desirable score, they only had one chance to take the EPT test on arrival at the ISU. In addition, the biggest difference in the setting of these two tests is the manner of test delivery and administration. While the TOEFL iBT is mediated by the Internet and computers, the ISU EPT test employs the paper-based format. For example, in the Listening section, instead of listening to the outside loudspeakers in a big hall as in the ISU EPT test, test-takers wear phones to listen and again use a mouse to choose the answer. These differences might create an impact on the performances of test-takers on these two tests.

Similar comparisons can be found as examining the test rubrics of these tests including instructions, test structure and time allotment. The instructions for both these tests are in the target language (English) and is in both aural and visual channels. Regarding test structure, some major differences and similarities are highlighted. The TOEFL iBT test consists of four sections, listening (20-40 minutes), speaking (20 minutes), reading (60-100 minutes) and writing (50 minutes). The reading section of the TOEFL iBT is comprised of three to five passages with approximately 12 to 14 questions each, totaling between 36 and 70 questions. The listening section of the TOEFL iBT, according to the official TOEFL website, contains four to six lectures (with or without classroom discussion) of approximately three to five minutes each with six questions each; two to three conversations of about three minutes long with five questions each. The total listening section contains between 34 and 51 questions. A third section is the speaking component, where test takers have to express their opinion on a familiar topic in two tasks and speak on what they read or listened to in four tasks, thus in total, completing six speaking tasks. The final section of the TOEFL iBT test is that of a writing component. Here, test takers are required to complete two writing tasks, one based on what they read or listened to and one in support of an opinion on a given topic (TOEFL iBT website). The EPT test at ISU consists of three sections: writing (60 minutes), reading (40 minutes) and listening (60 minutes). The EPT Listening test usually follows the other two sections. Both the TOEFL iBT and the EPT test are

of a fixed sequence, with clearly distinguished sections, and all the sections are of equal importance.

Some features in the scoring method and score scales between the two tests should be noted as well. In terms of scoring method, the scores of the Listening and Reading sections in both the ISU EPT test and the TOEFL iBT are based on the number of correct answers. Although the possible score of each section of the two tests is 30, there is a weighting among different test items in these two sections of the TOEFL iBT test. Moreover, the Speaking and Writing sections of the TOEFL iBT are scored holistically from a range of 0-4 for Speaking and 0-5 for Writing, and are then converted to a scale of 0-30. Meanwhile, the Writing section of the ISU EPT test is graded holistically using letters B, C, or D standing for one of the writing classes. Accordingly, the possible total score of the TOEFL iBT test is 120 points while the result on each section of the ISU EPT test is separately reported for each test taker.

Another noticeable difference in the constructs measured by these two tests is the integration of integrative skills and speaking skills in the TOEFL iBT test which are not tested in the ISU Placement test. For example, test-takers have to both read a short excerpt and listen to a short lecture to answer a question in either written or spoken form.

TOEFL iBT Listening vs. EPT Listening test

Based on the specification for listening measures of the TOEFL iBT (Chapelle et al., 2008), and the EPT Listening test specification (2007) with the examination of the EPT Listening test booklet used in the Fall 2010 administration (Set C2), more insightful observations about similarities and differences between the listening section of the two tests can be provided which are summarized in Table 17 below.

Table 18: Summary of the comparison of the specification for the TOEFL iBT listening measures (Chapelle et al., 2008, p. 193 & p. 243) and the EPT Listening Fall 2010 test booklet (Set C2)

Listening Claim	***Test takers can understand spoken English in an academic environment***			
Subclaims (Language abilities)	TOEFL iBT & EPT Listening	Basic understanding: understand the overall gist, important points and supporting details of lectures and conversations	Pragmatic understanding: understand the speaker's purpose for making a statement in a lecture or a conversation; understand the speaker's stance either the attitude expressed or the degree of certainty.	Connecting information: can understand connections between or among pieces of information in a single stimulus; can integrate information, draw inferences and conclusions, form generalizations, and make predictions on the basis of information heard in lectures and campus-based conversations
Nature of listening task	TOEFL iBT & EPT Listening	Questions about main ideas and important supporting details	Questions about a speaker's attitude or purpose, a speaker's degree of certainty, or a speaker's source of information	Questions about the relationships among ideas or about the organization of the aural text
Response types	TOEFL iBT	Simple selected response	Simle or complex selected response	Simple and complex selected response
	EPT Listening	Simple selected response	Simple selected response	Simple selected response
Scoring rubric	TOEFL iBT	Dichotomous right/wrong (0-1)	Dichotomous right/wrong 0-1; partial credit 0-2	Dichotomous right/wrong 0-1; partial credit 0-2
	EPT Listening	Dichotomous right/wrong (0-1)	Dichotomous right/wrong (0-1)	Dichotomous right/wrong (0-1)
Number of questions	TOEFL iBT (total 34	16-21; at least 6 main ideas	6 to 10 questions about speakers' purpose and attitude	6 to 10 questions about the relationships among ideas and about the organization of the text
	EPT Listening	17	8	5
Total time	TOEFL iBT	Approximately 50 minutes for 34 questions		
	EPT Listening	Approximately 60 minutes for 30 questions		
Nature of stimulus material	TOEFL iBT	All texts have a lead-in by a narrator and at least one context visual. Some have content visuals as well. *2 conversations:* one conversation in each form is in the office setting and includes interaction between a professor and a student (may include academic content); the other conversation is a service encounter (interactions between a student and a nonstudent that take place in university-related setting and have nonacademic content). (2-3 minutes) *4 lectures:* The content of the lectures is representative of an introductory level academic lecture; they present a variety of academic subject matter. Lectures may be monologic or interactive. In the interactive lectures, a student may ask the professor a question; the professor may ask the student a question and someone responds; and a student may comment on what the professor has said. Typically, half of the lectures in a form are interactive. (3-5 minutes)		
	EPT Listening	All the aural texts have a lead-in by a narrator and have videos. *4 lectures:* The content of the lectures is representative of an introductory level academic lecture; they present a variety of academic subject matters. Three out of the four lectures in Set C2 are found to be monologic. The other lecture includes only one comment from the second speaker; hence, it cannot be seen to be interactive. (2:33-7 minutes)		

As comparing the Listening sections of the two tests by using their test specification and the EPT Fall 2010 test booklet (Set C2), some different and similar features are noted in terms of the language of input. First, while the TOEFL iBT Listening test clarifies two academic settings for testing listening which are on-campus conversations, and lectures with or without the participations of students, the input of the Listening section in the ISU EPT test consists of only authentic lectures. Second, instead of using time to describe the Listening input as in the TOEFL iBT specification (2008), the input for this section of the ISU EPT test is based on the number of words in the recording or the length of academic sources, i.e short speech samples and extended academic listening passages. While the short sample consists of only 10 to 12 words, the long lecture mostly has up to 600 words. In addition, the two sections of the tests are totally different in delivery and speededness. For instance, the ISU Listening Placement test is a paper and pencil based test, test-takers can return to previous questions and change their answers on the answer sheet; on the contrary, the Listening Section in the TOEFL iBT test does not allow test-takers to return and change their answers. A noticeable similarity in this section between the two tests is that they allow test-takers to take notes if necessary. Their input both include visual aids such as graphs, charts to support the listening by providing contexts for test-takers. However, the input of the Listening section in the ISU EPT test contains video which is expected to require different metacognitive skills of the test takers to complete its test tasks.

Next, the examination on question format and expected responses in the TOEFL iBT Listening test, the EPT Listening test reveals that question types and formats in the TOEFL iBT Listening test are more diverse than those in the EPT Listening test despite similar forms of expected responses. In specific, the Listening sections of the two tests both mainly use the multiple choice format, and their expected responses are partially structured. However, their diversity in question types is different between them. For example, the ISU EPT Listening test items seem to be simpler as all the questions have four options provided whereby one choice represents the correct answer, one choice is plausible (not incorrect given the context), one choice is too narrow, and one choice is too broad. Nevertheless, the questions in the TOEFL iBT Listening test are more various with up to four different formats. They include traditional multiple-choice questions with four answer choices and a single correct answer, multiple-choice questions with more than one answer (e.g., two answers out of four or more choices), questions

that require test takers to order events or steps in a process, and questions that require test takers to match objects or text to categories in a chart (ETS, 2011)

Noticeably, in terms of constructs to be measured, both similarities and differences in the Listening section between the two tests can be seen in terms of language knowledge, competence of test-takers, and strategies. In the similar way, the Listening sections of the two tests are aimed at measuring how well test-takers understand spoken English language in the academic environment. Furthermore, the task-types to measure the constructs in the Listening section in these two tests are totally the same including listening for basic comprehension, listening for pragmatic understanding, and listening for synthesizing information (ETS, 2011; EPT Listening Specification, 2007). For instance, regarding the construct of listening for pragmatic understanding, both tests have multiple-choice questions about the understanding of a speaker's attitude, degree of certainty, or purpose in the Listening section. These questions require test takers to listen for voice tones and other cues, and determine how speakers feel about the topic they are discussing. The last not the least, the specifications used for the listening section of the TOEFL iBT and the EPT tests are claimed to base on the model presented by Davidson and Lynch (2002).

The reasons why such extended descriptions of the tests are provided are for the purpose of allowing me to make inferences regarding the various theoretical constructs they assess. Bachman (2004) explains that although correlations can be calculated to statistically show the relationship between two variables, the relationship is made meaningful if it is interpreted in view of its application (p. 79).

(b) Hypothesis

As looking into the descriptions of the two versions of the TOEFL test and the EPT Listening test at ISU, some questions about the relationships among these tests arose. Firstly, what is the relationship between the TOEFL iBT Listening scores and the EPT Listening scores in Fall 2010 at ISU? Do they measure the same thing or not? Then, at the same time, how does the TOEFL iBT total score set correlate with the EPT Listening score set? Secondly, did the students who performed well on the TOEFL pBT test, have a good performance on the EPT Listening score set as well? These questions lead to some hypotheses on their correlations and motivate me to carry out statistically correlation analyses in order to examine these claimed hypotheses.

Based on the comparative review about the three tests above, moderately positive correlations among the four pairs of test score sets are expected with the coefficients higher than 0.5. However, the strengths among them probably vary. In specific, because of a number of shared constructs in the Listening section of the two tests (TOEFL iBT and EPT Listening test), their two score sets of the TOEFL iBT Listening test and the EPT Listening test will co-vary greater than the two score sets of the TOEFL iBT test and the EPT Listening test. Also, the relationship between the TOEFL pBT total scores and the EPT Listening test scores will be stronger than that between the TOEFL iBT total scores and the EPT Listening scores. Accordingly, those whose TOEFL iBT Listening scores are higher will be more likely to have higher scores on the EPT Listening Fall 2010 test; nevertheless, those whose TOEFL iBT total scores are higher might not perform better on the EPT Listening test. Thus, although the three tests appear to be theoretically strongly correlated, we can assume that with the consideration of the error factors, the correlation coefficients would probably fall in the region from 0.6 to 0.8 for the score sets of the three tests.

(c) Results

As mentioned earlier, the dataset for the correlation investigation is comprised of scores of 395 out of 556 test-takers in the EPT Fall 2010 administration at ISU. However, due to the offering of two different versions of the TOEFL tests, and the missing of the TOEFL listening component scores, the number of test scores for each set of the TOEFL test varies. Therefore, four designated correlation analyses of the test-takers of the EPT Listening Fall 2010 administration will be based on four different pairs of score sets including: (1) between the EPT Listening Fall 2010 score set and the TOEFL pBT score set, (2) between the EPT Listening Fall 2010 score set and the TOEFL iBT Listening score set (3) between the EPT Listening Fall 2010 score set and the TOEFL iBT total score set, (4) between the EPT Listening Fall 2010 score set and the TOEFL iBT total score set including the converted TOEFL pBT scores.

(c1) Descriptive statistics

Table 18 displays some main general statistical results of the four pairs of data sets highlighting the number of students as well as the mean, median, mode, standard deviation and other descriptive indexes for distribution of each score set. In order to better compare the variation in these indexes in different tests using different score scales, the results were also

converted into the percentage out of the total possible score. The results of the conversion are presented in the table as well.

Table 19: Summary of descriptive statistics of four pairs of score sets of the test-takers of the EPT Listening Fall 2010 administration at ISU

Tests	TOEFL pBT		TOEFL iBT Listening		TOEFL iBT total score		TOEFL iBT total scores with TOEFL pBT converted scores	
Pairs	EPT Listening	TOEFL pBT	EPT Listening	TOEFL iBT Listening	EPT Listening	TOEFL iBT	EPT Listening	TOEFL iBT
Mean	13.92	508.08	16.47	21.17	16.55	85.86	16.21	83.17
%	*0.46*	*0.77*	*0.55*	*0.71*	*0.55*	*0.72*	*0.54*	*0.69*
Standard Error	0.58	8.44	0.28	0.28	0.24	0.59	0.22	0.73
%	*0.02*	*0.01*	*0.01*	*0.01*	*0.01*	*0.00*	*0.01*	*0.01*
Median	14	520	17	21	17	86	16	84
%	*0.47*	*0.79*	*0.57*	*0.70*	*0.57*	*0.72*	*0.53*	*0.70*
Mode	14	543	19	24	15	80	15	80
%	*0.47*	*0.82*	*0.63*	*0.80*	*0.50*	*0.67*	*0.50*	*0.67*
Standard Deviation	4.16	60.26	4.46	4.45	4.39	10.91	4.44	14.42
%	*0.14*	*0.09*	*0.15*	*0.15*	*0.15*	*0.09*	*0.15*	*0.12*
Min	6	393	4	7	4	38	4	27
Max	25	633	27	30	27	110	27	110
Kurtosis	0.48	-0.91	-0.19	-0.28	-0.11	0.76	-0.30	1.71
Skewness	0.71	0.04	-0.27	-0.35	-0.31	-0.48	-0.19	-1.09
Range	19	240	23	23	23	72	23	83
%	*0.63*	*0.36*	*0.77*	*0.77*	*0.77*	*0.60*	*0.77*	*0.69*
Number of test-takers	51	51	258	258	344	344	395	395

Some main observations about their general statistics can be reported here. First, the large majority of 395 test-takers of the ISU EPT Fall 2010 took the TOEFL iBT test in comparison to a small number of test scores in the TOEFL pBT test score set (N=51). A look at the results of the TOEFL pBT score sets also reveals the existence of some extremes. For example, out of the eight score sets, the TOEFL pBT score set took the first position in mean (0.77%), medium (0.79%), mode (0.82%), but had the lowest indices in terms of range (0.36%), standard deviation (0.09%), and standard error (0.01).

Secondly, the TOEFL score sets were found to comprise of more high scores while the EPT Listening score sets had more low scores. As can be seen, three indices (mean, mode, median) of the EPT Fall 2010 Listening test score set were the lowest, equal to half of its possible total score. On the other hand, the results of these three indices in the TOEFL-related score sets were all pretty higher than half of their possible total scores.

Next, for the TOEFL iBT related score sets, there are few disparities in the descriptive statistical results between the TOEFL iBT Listening score set and the two TOEFL iBT total score sets. For example, their mean scores ranged from 69% to 72 % out of their possible scores. Similarly, their median results made up for 70% to 72% out of their possible scores although the mode score of the TOEFL iBT Listening score set (80% out of the possible score) was found to be much higher than the other TOEFL iBT score sets (67% out of the possible score).

The third surprise is with the score ranges of the TOEFL pBT total score set (393-633), and the TOEFL iBT score set (38-110). In specific, 24 out of 51 test takers under examination who were admitted into the ISU were found to have scores lower than 519. Likewise, more than 14 test-takers whose TOEFL iBT total scores were lower than 72 were counted in the given score set.

Finally, the two Listening score sets (EPT Listening and TOEFL iBT Listening) had a less variation among these indices in comparison to other pairs of score sets. For example, in terms of standard deviation, on the average the scores in these two sets both equally deviated from the mean of the score set (0.15).

(c2) Score distribution

As can be seen from Table 18, all the score sets had indices of skewness and peakness ranging from -2 to 2 which suggests that they all had acceptable distributions. The distributions of the four pairs of score sets are also visually presented in figures 6 to 9 below.

The distributions of the two score sets on the EPT Listening Fall 2010 test and the TOEFL pBT test are presented in Figures 6-A and 6-B. As discussed above, these score sets had the smallest number of test takers (N=51) which will certainly influence on their reliability, and distribution. Accordingly, the EPT Listening score set varied from 6 to 25 out of a possible 30 whereas the TOEFL pBT score set ranged from 393 to 633 out of a possible 660. The positive skewness indices of these two score sets (0.71 and 0.04) indicated that the scores on the two tests (EPT Listening and TOEFL pBT) were positively skewed. Also, the scores were seen to move

towards to the left of the center line of the graph. Thus, there were more scores lower than the mean score in these two score sets. The greater value of the kurtosis index of the TOEFL pBT score set suggested that the distribution of this score set was more peaked than that of the EPT Listening score set. Meanwhile, its negative value showed that more higher scores out of the possible total score were found in the TOEFL pBT score set than the EPT Listening test and vice versa. Noticeably, a much higher percentage of the test scores in the EPT Listening data set (0.80) fell within one standard deviation from the mean score whereas a much lower percentage of the test scores in the TOEFL pBT data set (0.51) was within a similar range. As a result, the distribution of the designated EPT Listening test score set in Fall 2010 was found to be more normal than that of the TOEFL pBT score set of 51 EPT test-takers.

Figure 6-A: Distribution of the EPT Listening Fall 2010 score set of the test takers with TOEFL pBT scores (N=51)

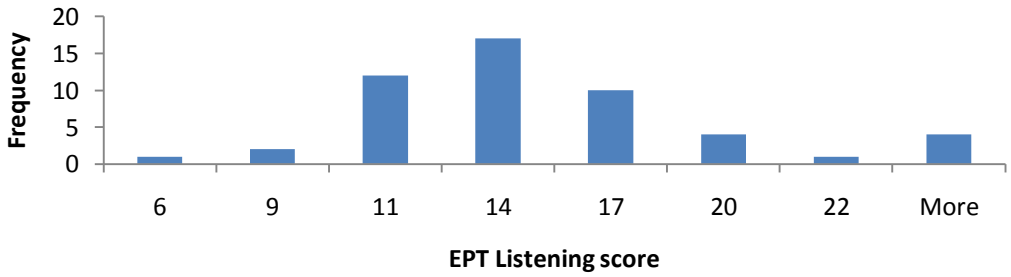
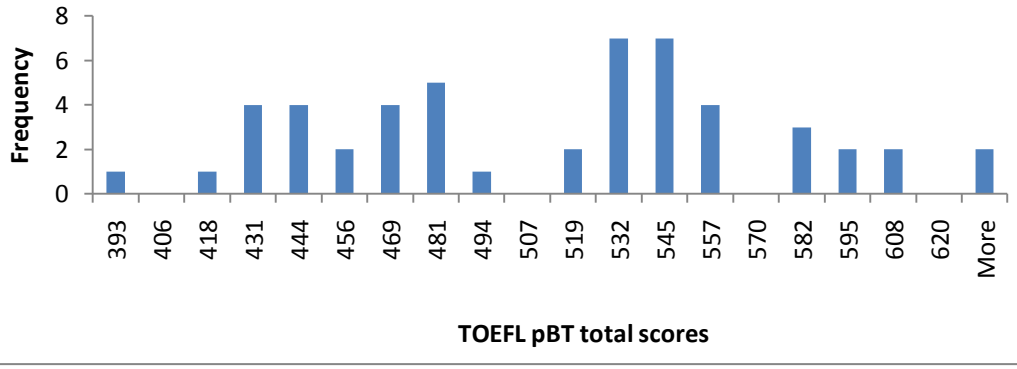


Figure 6-B: Distribution of the TOEFL pBT score set of the EPT Fall 2010 test takers at ISU (N=51)



Figures 7-A and 7-B illustrate the distributions of the EPT Listening test score and the TOEFL iBT Listening test score of 258 test takers of the ISU EPT Fall 2010 test. Both these tests have the same total possible score of 30. Although both of the score sets had the same range (23), their two extreme scores were different. The lowest score and the highest score obtained on the

EPT Listening Fall 2010 test were 4 and 27 out of a possible 30 while those on the TOEFL iBT Listening test were 7 and 30 out of a possible 30 respectively. However, we can further see that the scores obtained from both tests under examination had a pretty normal distribution. Even though their percentages within each standard deviation were not exactly such as those typically described by a normal distribution, these percentages were closely related, with the greatest percentage of students scoring within one standard deviation of the mean (61% for the EPT Listening test score set, and 68% for the TOEFL iBT Listening test score set). The histograms also showed that the two score sets were negatively skewed as the scores were observed to move towards the right side of the mean score or the center line of the graph. Moreover, the score distributions were detected to be rather flat (-0.19 and -0.28). In other words, there were not many too high or too low scores in the two sets.

Figure 7-A: Distribution of the EPT Listening Fall 2010 score set of the test takers with TOEFL iBT Listening scores (n=258)

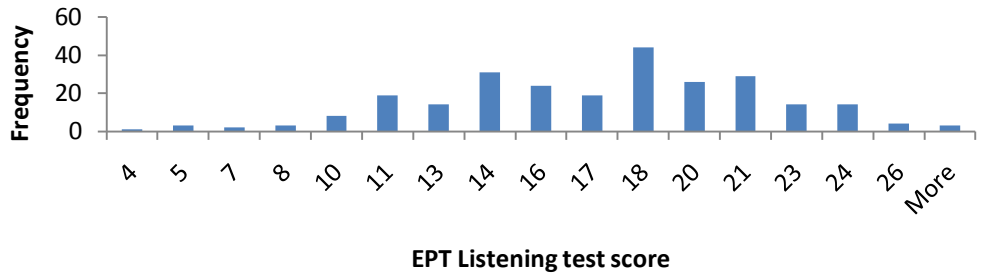
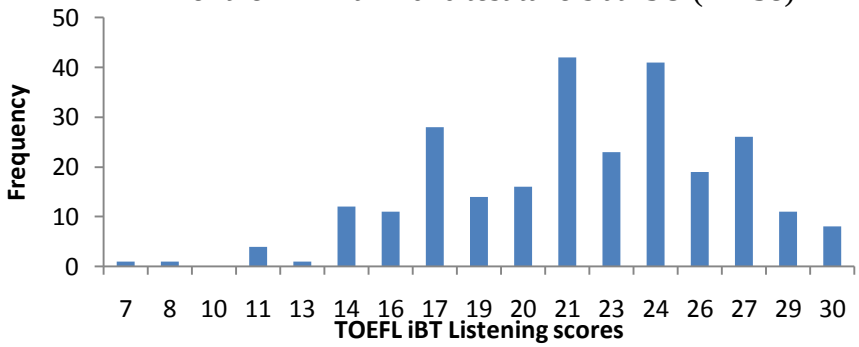


Figure 7-B: Distribution of the TOEFL iBT Listening score set of the EPT Fall 2010 test takers at ISU (n=258)



Figures 8-A and 8-B show the distributions of the score sets on the EPT Listening test and the TOEFL iBT test. In comparison to the previously reported listening score sets on the two

tests, there were additional 86 test-takers who had the TOEFL iBT total score, but missed its listening score. As can be seen, the EPT Listening Fall 2010 score set ranged from 4 to 27 out of 30 while that of the TOEFL iBT varied from 38 to 110 out of 120. In other words, the score range of the EPT Listening Fall 2010 score set was wider than that of the TOEFL iBT score set. This may explain the reason why the EPT Listening score set was seen to be better normally distributed than the TOEFL iBT total score set. Noticeably, 64% of the EPT Listening Fall 2010 test scores fell within one deviation from the mean score while more than 68% of the TOEFL iBT total scores belonged to this group. The negative skewness indices of these two score sets also showed that the majority of the test-takers under examination scored higher than the mean score on both tests. Despite their acceptable kurtosis indices, the kurtosis of the TOEFL iBT score set which was positive and greater in value (0.76 vs. -0.11), revealed that the number of test-takers who had noticeably higher scores was much greater on the TOEFL iBT test than on the EPT Listening Fall 2010 test.

Figure 8-A: Distribution of the EPT Listening Fall 2010 score set of the test takers with TOEFL iBT total scores (N=344)

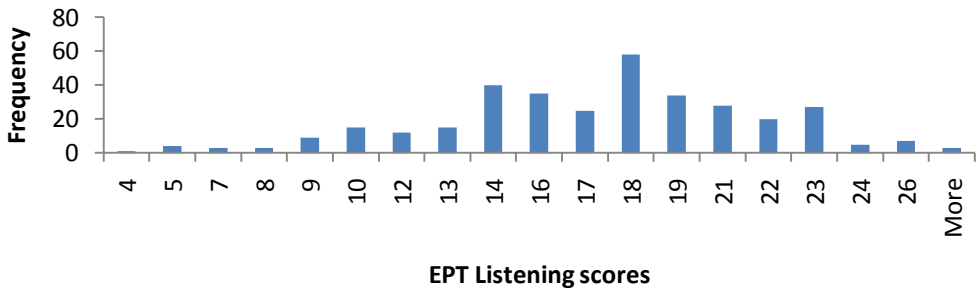
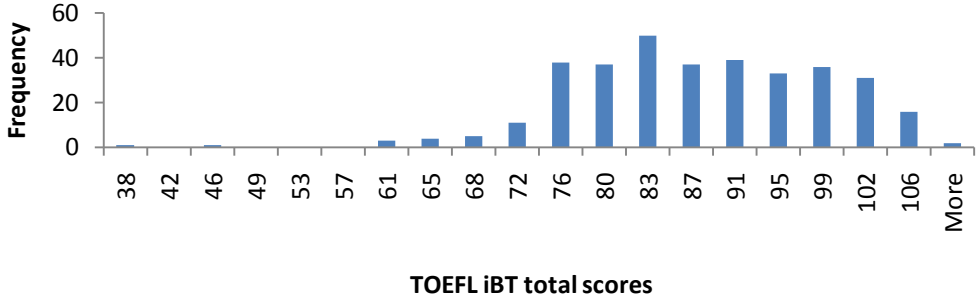


Figure 8-B: Distribution of the TOEFL iBT total score set of the EPT Fall 2010 test takers at ISU (N=344)



The final score distributions under examination were based on the EPT Listening Fall 2010 score set and the TOEFL iBT total score set with the converted TOEFL pBT test scores of 395 test-takers. They are presented in Figures 5-A, and 5-B respectively. As can be seen, the EPT Listening Fall 2010 test score set had a low score of 4 and a high score of 27 whereas the TOEFL iBT total score set varied from 7 to 110. In other words, these two score sets covered a wide score range.

In fact, both Figures 9-A and 9-B highlight the variance in scores obtained, and the TOEFL iBT score set for this particular group of students showed a worse distribution than the EPT Listening score set. Noticeably, while 64% of the EPT Listening scores belonged to the range within one deviation from the mean score, up to 74% of the TOEFL total scores were within this range. Morever, their negative skewness indices suggested that the scores in the two score sets moved towards to the right of the center line. Also, the value of skewness of the TOEFL iBT total score set was much higher than that of the EPT Listening test. In other words, a greater number of scores which were higher than the mean score was found in the TOEFL iBT score set than in the EPT Listening score set. In addition, the positive and high kurtosis index of the TOEFL iBT total score set (1.76) indicated that the score set was very peaked with a lot of extreme scores at the lower end of the score range. On the contrary, the more reasonable kurtosis value (-0.3) of the EPT Listening test score showed that the score distribution was rather flat with few extreme scores.

Figure 9-A: Distribution of the EPT Listening Fall 2010 score set of the test-takers with TOEFL scores (n=395)

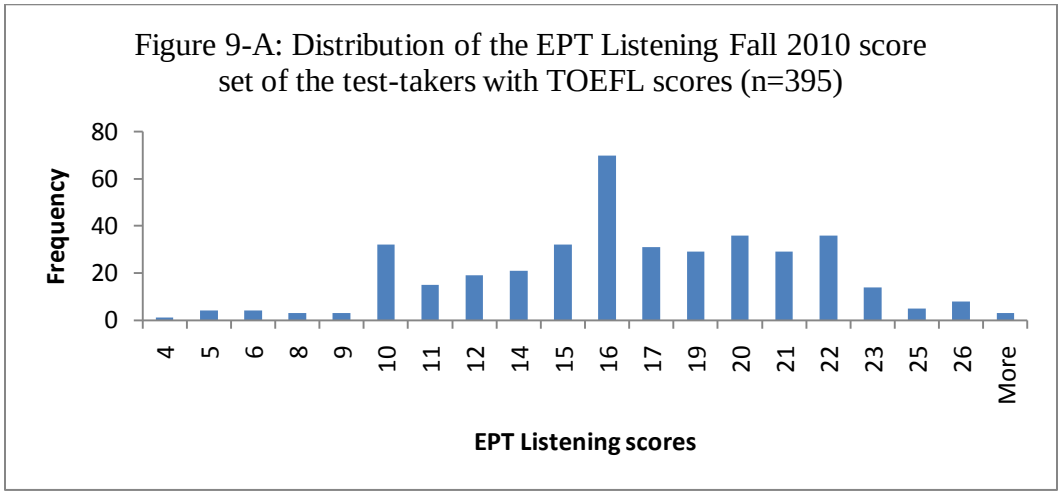
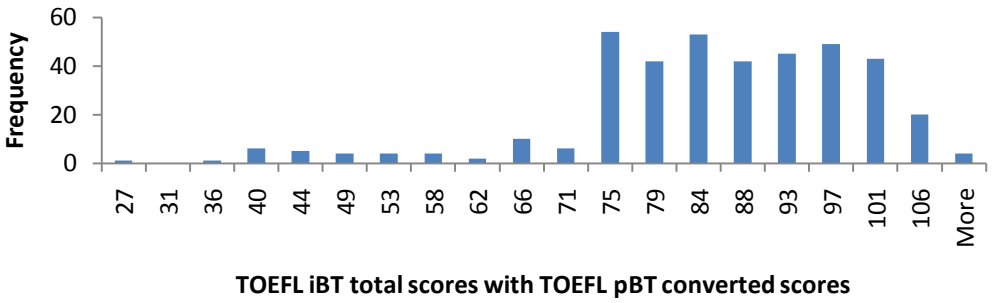


Figure 9-B: Distribution of the TOEFL iBT converted score set of the EPT Fall 2010 test-takers at ISU (N=395)



(c4) Correlation results

The report of the results of four correlation analyses among different score sets of the test-takers of the EPT Listening Fall 2010 administration starts with an examination into the nature of their relationships leading to a decision on the use and production of a particular correlation coefficient as a statistical evidence on their correlation relationships.

The relationships of the four pairs of score sets are visually represented through the four scatterplots (Figures 10-13) below. As can be seen, these test score sets had linear patterns as an imaginary line could be drawn along these spots. These relationships portrayed in the scatterplots seem to be both positive, however not necessarily strong.

Their positive relationships with varying strengths among these four pairs of the test scores can be better explained by a close look at the density of spots distributed in these scatterplots. Accordingly, the score sets of the TOEFL iBT total scores with and without the TOEFL pBT converted scores and the EPT Listening test (Figures 13) appeared to have stronger relationships because the spots were the most concentrated to form a linear line despite several outliners. Also, the scatterplot of the score sets of the TOEFL iBT test and the EPT Listening test by 344 test-takers at ISU (see Figure 12) was observed to have a greater density of spots, but there existed quite a few outlines. Furthermore, as looking at the scatterplot of the TOEFL iBT Listening scores and the EPT Listening test scores (see Figure 11), the spots were seen to be more dispersed. Therefore, the relationship between the two data sets were expected to be weaker than that of the other three pairs of data sets. Finally, for the scatterplot of the TOEFL pBT and the EPT Listening test (see Figure 10), the small number of spots or scores might cast

doubts on the reliability of the strength of its relationship even though the two score sets were seen to have the best correlation.

Figure 10**:** The relationship between the students' performance on the TOEFL pBT and on the EPT Listening test in Fall 2010 at ISU

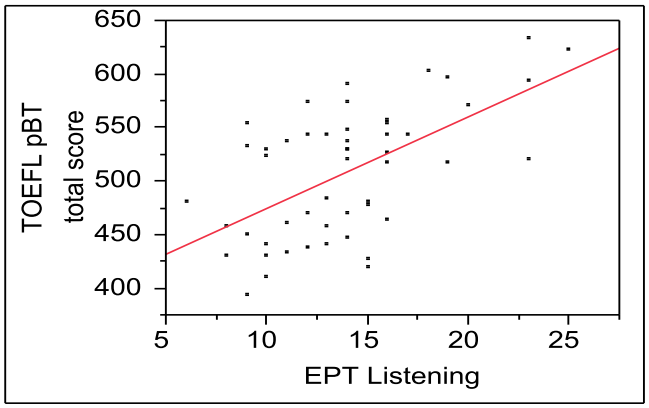


Figure 11: The relationship between the students' performances on the TOEFL iBT Listening test and on the EPT Listening test in Fall 2010 at ISU

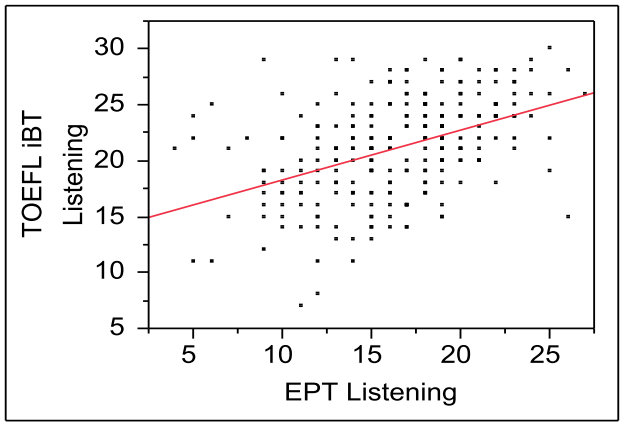


Figure 12: The relationship between the students' performances on the TOEFL iBT and on the EPT Listening test in Fall 2010 at ISU

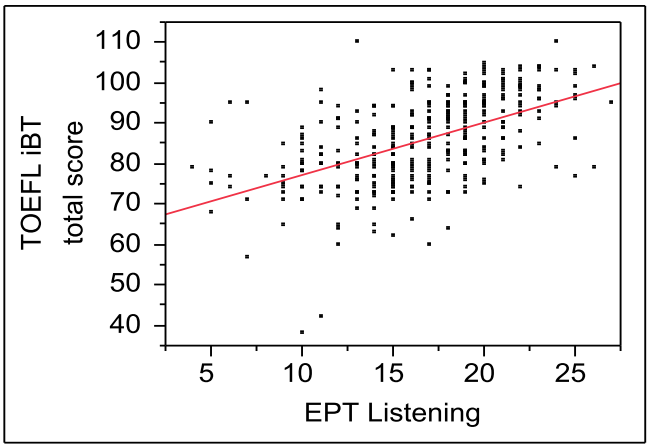
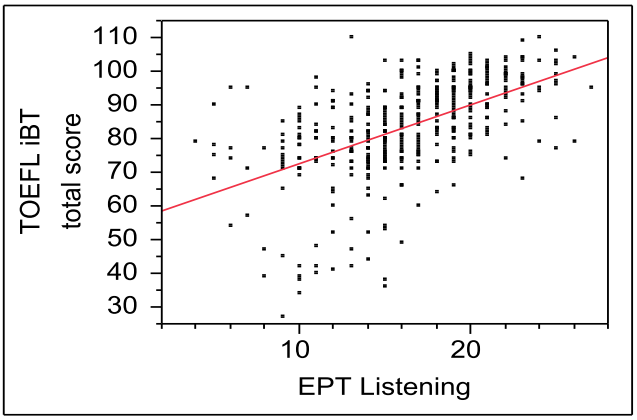


Figure 13: The relationship between the students' performances on the TOEFL tests using the TOEFL iBT score scale and on the EPT Listening test in Fall 2010 at ISU



Due to the ambiguity of the visual representation of the scatterplots for the four pairs of score sets, it is necessary to employ a statistical index – correlation coefficient in order to determine the actual strength between each pair. As being described above, the four pairs of score sets all meet the three conditions to use the Pearson product-moment correlation coefficient formula (Bachman, 2004, p. 85). In specific, the relationships among any two of these test score sets are linear, and these scores constitute interval scales as well as are pretty normally distributed. Thus, the Pearson product-moment (r) is chosen in this investigation of correlation among the designated score sets. Table 19 below displays the results of the correlation

calculations for the four pairs of score sets. They will then be reported in terms of strength, direction and statistical significance.

Table 20: Summary of Pearson product-moment correlation coefficients for four pairs of score sets by the test-takers of the EPT Listening Fall 2010 administration at ISU

Variable	**By variable**	**Correlation *(p<0.01)***	**Count (N)**
TOEFL pBT	EPT Listening	0.5867	51
TOEFL iBT Listening	EPT Listening	0.4477	258
TOEFL iBT total score	EPT Listening	0.5228	344
TOEFL iBT total score (with converted TOEFL pBT score)	EPT Listening	0.5387	395

In accordance with the observed relationships from the above scatter plots and the predictions based on their descriptive statistics, all the score sets under examination positively but not strongly correlate with each other. Accordingly, the TOEFL pBT score set and the EPT Listening score set had the highest covariance coefficient of around 0.587 while the TOEFL iBT Listening score set and the EPT Listening score set owned the lowest correlation index of approximately 0.448. The other two pairs of score sets (TOEFL total scores, and EPT Listening scores) had a nearly similar correlation strength from 0.523 to 0.539.

In terms of statistical significance, all correlation coefficient results are compared with the critical values of correlation coefficients for the corresponding number of test scores in a certain score set at the highest probability level of 0.01. Accordingly, the critical value for a group of 60 items to ensure its 0.01 probability to not occur is 0.325, and this value decreases as the number of items in the group increases. As a result, it is found that all the correlation coefficients obtained from the four pairs of score sets by the EPT Listening test-takers in Fall 2010 at ISU, are statistically significant and far higher than the critical values for the set probability of 0.01. In other words, we can be 99% confident about the existence of their relationships.

In short, the examination on the strength, direction, and significance of the collected correlations among the four selected pairs of score sets supports that we can trust the positive covariance of the designated data sets by the EPT Listening Fall 2010 test takers at ISU. However, their strengths of correlation are quite low and need improving. An insight into the meaningfulness of the correlation results will be given in the next part.

(d) Discussion

The results of the statistical analysis show both some consistencies and inconsistencies with the original expectations discussed in the hypothesis in terms of direction, strength, and statistical significance. First, the correlation coefficients support the existence of the positive relationships among the designated test score sets of the test takers of the EPT Listening Fall 2010 administration despite their varying strengths. In other words, the observed correlations among the score sets under examination did not happen by chance. Also, as being expected, the correlation coefficient of the EPT Listening scores and the TOEFL iBT total scores with the TOEFL pBT converted scores (approximately 0.54) was a bit higher than that of the EPT Listening scores and the TOEFL iBT total scores without the TOEFL pBT converted scores (about 0.52).

However, some differences can be highlighted. While it was suspected that the correlation coefficient between the EPT Fall 2010 Listening test score set and the TOEFL iBT test score set would be the highest, the correlation coefficient between these two tests (approximately 0.45) was shown to be the lowest out of the four score sets. On the other hand, contrary to my expectation about its moderate correlation, the score set of the test takers who took the TOEFL pBT test was found to most correlate with their EPT listening test score set.

The lower correlation results than expected might be contributed by a number of both internal and external factors. According to Bachman (2004), four possible common causes to the problems with the production of a correlation coefficient in language testing and assessment include measurement error, distributional characteristics of the data under examination, the existence of extreme cases, and a combination of multiple groups. As looking into the four pairs of the score sets of the EPT Listening Fall 2010 test-takers and taking Bachman's suggested possible main causes into consideration, some explanations about their low correlation coefficients can be presented here.

First, there are numerous internal and external factors contributing to the error variance in the two tests. In specific, for the Listening section in the ISU EPT test, the result of the EPT test analysis show that the test booklet needs improving due to its low reliability index, as well as its low item difficulty and discrimination indices. For external factors, the test takers might have had unfavorable physical conditions such as tiredness, or illness due to jet lag, and the differences in their performances on the test might have been attributed by their familiarity with

test formats. Other internal factors influencing the test reliability of the TOEFL iBT test scores can, for example, be rater's consistency, topical or content knowledge while its external factors might be test-takers' familiarity with computers, the Internet, and question types.

Secondly, as discussing about the distributional characteristics of each pair of score sets by the EPT Fall 2010 test takers, although the score sets are proved to be linear, and normally distributed, some possible issues are observed. As mentioned above, most of the students' test scores under investigation in this paper are students who are admitted to Iowa State University, they generally have met the TOEFL iBT requirement of at least 71 of 120 (equivalent to 530 TOEFL pBT). This range restriction on the test scores of only the admitted international students at the ISU will probably affect the distribution of its test scores, as well as its reliability because the TOEFL tests are aimed at a larger group of audiences with much wider range of language proficiency. There also exist a number of outliners or extreme cases in the TOEFL iBT score sets which are too low in comparison to others in the score sets. In short, the truncated samples in the TOEFL score sets and the existence of extreme cases might have affected the strength of the yielded correlation coefficients among the designated score sets of the EPT Fall 2010 test-takers.

Next, Bachman (2004, p. 98) explains that combining multiple groups can either increase or decrease the strength of correlation between them. This claim might also contribute to the reason why the correlation coefficient between the EPT Listening score set and the total TOEFL score set by the test-takers of the EPT Fall 2010 administration was observed to be higher than that between the Listening sections in the two tests (EPT and TOEFL iBT). In this case, the TOEFL total score comprises of four component scores three of which might make up for the listening score. Likewise, the conversion of the TOEFL pBT scores into the TOEFL iBT total scores might have created some influence on the relevant final correlation result.

2. Construction of the validity argument for the EPT Listening Fall 2010 test at ISU

The expected outcome of this section is to construct a validity argument for the EPT Listening Fall 2010 test score interpretation and use. However, it should also be noticed that all the inferences with their warrants in the validity argument cannot be addressed within a single study. In other words, it requires an accumulation of studies to produce backings for or against all the assumptions of relevant warrants for each inference. The following discussion presents a partial construction of the validity argument for the EPT Listening Fall 2010 at ISU in response

to the research questions, which follow the order of inferences in the proposed validity argument structure presented in Chapter 2 (see Table 3).

How do the EPT Listening test design and development help to measure what we want to measure of test-takers? (Warrants 1 & 2)

As being seen in the proposed validity argument for the EPT Listening test in Fall 2010, the domain definition inference is based on the warrant that observations of performances on the EPT Listening test reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education, especially in mid-western areas of the U.S.A. This warrant, in turn, is based on the two assumptions including (1) critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified, (2) assessment tasks that require important listening sub-skills and are representative of the academic domain can be simulated. The results of the qualitative examination on the EPT Listening test specification and the test booklet used in the Fall 2010 administration (Set C2) show that the test design and development give both positive and negative attributes to the two assumptions.

For the first assumption, the specification for the EPT test is claimed to base on the framework of academic skills and sub-skills provided by Bachman. A number of important academic skills are identified such as listening comprehension, listening for main ideas, and listening for inferences. However, the definition of academic listening constructs to be measured in the EPT test is found to be ambiguous and general. In fact, the descriptions of academic listening skills and sub-skills in the specification fail to show how the academic listening skills measured by the EPT test are different from other measurements, especially the TOEFL test. In addition, few theoretical and empirical evidences are provided to lay the basis for such measurement in the test. Therefore, it is difficult to see how well the design and development of the test based on its specification serve the distinctive purpose of the EPT test described in the test manual.

For the second assumption, the analysis of the test set (Set C2) also signifies both strengths and weaknesses. The test was found to be very authentic for using real lectures in the U.S. In specific, these lectures include video and cover a wide range of academic topics in various accents. Moreover, three out of the four listening texts were developed carefully following main steps of the test development, namely developing, testing, and revising.

Likewise, most of the questions in the test set were evaluated to measure the test-takers' academic listening skills and relevant sub-skills such as comprehension, or inference. Thus, on the whole, the design and development of the test set (Set C2) is evaluated to have a strong emphasis on real-life academic listening which is found to meet the purpose of the EPT test, and to be suitable for the specific group of admitted international students at ISU. Nevertheless, the last lecture in the test set was detected not to measure the test-takers' academic listening skills and sub-skills well due to several bad test items. Some limitations on the diversity of question types, formats, scoring and instructions in the test set were also notified which might have restricted the simulation of real-life tasks in the academic domain.

How reliable is the EPT Listening test in measuring test-takers' proficiencies? (Warrant 3)

The generalization inference in the interpretative argument for the EPT Listening test relies on the warrant that observed EPT listening scores are estimates of expected scores over the relevant parallel versions of listening tasks and test forms and across raters. Then, the warrant is supported by four assumptions which are (1) a sufficient number of tasks are included in the EPT listening test to provide stable estimates of test takers' listening performances, (2) configuration of tasks on listening measure is appropriate for intended interpretation, (3) appropriate scaling and equating procedures for EPT listening test scores are used, (4) EPT listening tasks and test specifications are well defined so that parallel tasks and test forms are created.

A number of backing which are both for and against some of the four assumptions are found. For the first assumption, the history of the EPT Listening test shows that the number of test items in Set C2 (n=30) is the outcome of numerous reliability analyses on previous test booklets. The number of test items in a test is thus evaluated to be sufficient to measure the test-takers' academic listening proficiency at an acceptable level of reliability and stability within the allowed time constraint. Moreover, the test item analysis of Set C2 suggests that most of the test items in the test have reasonable reliability indices. As mentioned above, although the question format in the test is limited to the traditional multiple-choice format, the configuration of question types shows its equal emphasis on measuring two main aspects of academic listening, specifically comprehension and inference. The distribution among different question types of the test set (Set C2) is also shown to be reasonable to measure academic listening proficiencies and correspond to the suggested distribution in the test specification. Other empirical evidences based on the test scores of the test takers of the EPT Fall 2010 administration, which include

some reliability estimates and its score distribution, indicate that the test set (Set C2) owns a suitable reliability estimate but still needs revision in order to reach the desired reliability estimate of 0.8. Likewise, the distribution of the score set in the EPT Listening Fall administration is found to be quite normal.

However, some negative points can be provided here. The scaling and equating procedure for EPT listening test scores are generally found to be very simple. For example, all the test items are treated equally without any difference in weighting. Furthermore, the investigation into the test specification for the EPT Listening test shows that the specification gives some general information about the test structure, characteristics of the input, and question types. However, the specification is proved to be incomplete and need improvement for better guidance, especially for the development of equivalent test tasks, and test forms. Finally, the mismatches found in the test structure specified in the test specification and the real test booklet, signify the necessity to invest more efforts into finishing the blue print for the EPT Listening test.

How do students' scores on other test of language development (TOEFL) correlate with their scores on the EPT Listening test? (Warrant 4)

The explanation inference is based on the warrant that expected listening scores in the EPT Listening test are attributed to the construct of academic listening proficiency. In turn, five different assumptions to validate the warrant are (1) the linguistic knowledge, processes, and strategies required to successfully complete listening tasks vary across tasks in keeping with theoretical expectations; (2) task difficulty is systematically influenced by task characteristics; (3) performance on the EPT listening test relates to performance on other test-based measures of language proficiency as expected theoretically; (4) the internal structure of EPT listening test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components; (5) test performance on the EPT Listening test varies according to amount and quality of experience in learning English. The assumption (3) is addressed in this study through a correlation analysis between performances on the EPT Listening test and on the TOEFL test of the EPT Listening Fall 2010 test-takers.

The results of the correlation analyses give some noticeable insights into the explanation inference of the EPT test score use and interpretation. The statistical evidences are found to respond to the theory-based hypotheses on their existence of relationships although the strengths

of their correlations are not as high as expected. In specific, the EPT Listening test and the TOEFL test are statistically proven to measure some shared constructs of the same language proficiency. On the other hand, the measurements of performances on the listening section of the two tests (EPT vs. TOEFL iBT) are shown to have some co-variance which is less stronger than the other pairs.

The comparative reviews of relevant tests (EPT Listening, TOEFL pBT, and TOEFL iBT) and the discussion of the factors influencing these correlation strengths lead to two deductions. First, the EPT Listening test is necessary to reassess the academic listening proficiency of the international students admitted to the ISU. The description of the TOEFL total score sets indicates that there are a number of students whose TOEFL scores are much lower than the required score. This fact in the admission of new comers to the ISU necessitates the administration of an EPT Listening test in specific and an EPT test in general in order to reassess their academic language skills for a specific group of ESL learners. In addition, the fact that the TOEFL iBT total score set and the TOEFL Listening score set have lower standard deviation, and higher mean results, shows that these tests might not to be able to discriminate the group of 395 test takers effectively. Finally, the low correlation coefficients collected in the analyses suggest that the TOEFL total scores and the TOEFL listening scores might not be able to predict the academic listening skills and the language skills of the test-takers in the authentic academic context.

Second, while mentioning the good attribute of the positive but low correlation coefficients between the two tests (EPT Listening vs. TOEFL tests) to the necessity to operate an EPT test at ISU, it is also important to notice the call for improving their strengths due to a number of their shared underlying academic constructs. The expected correlation coefficients should range from 0.7 to 0.8. Therefore, with reference to some negative points in the first three warrants in the interpretative argument for the EPT Listening test, the correlation results really cast some doubts on the validity of the EPT Listening test scores in explaining about the test-takers' academic listening skills.

What are challenges to the validity argument of the EPT Listening Fall 2010 test at ISU?

While all the answers to the three research questions above are taken into consideration, some significant challenges to the reliability and validity of the EPT Listening test in Fall 2010 are highlighted. They will be presented here in order to give a proper justification about the

validity argument for the EPT Listening Fall 2010 test in specific, and suggest some innovations on the EPT Listening test in a long term.

The first challenge is the logistic and administrational constraints. In fact, the annual number of international students is quite high, and has been increasing with some changes in the admission policy about language requirement. Thus, the target test-taker populations for the EPT test might be different among different terms, which might have lead to some effects on the results in reliability, and score distribution. Another constraint is the limited sources for the test. The test is now in charge by Prof. Hegelheimer and a Ph.D candidate Chung who are occupied with a lot of work. Thus, this constraint on personnel might explain the reasons why more efforts should be invested on the test design and development. Also, the test scores on the EPT test are supposed to be available online one day after the test date. This time constraint accounts for the adoption of the multiple choice format for automatic scoring. Thus, the manners of test-delivery, scoring and reporting should be considered in order to reduce the logistic and administrational constraints.

The second challenge with this mid-stakes test is how to balance resources to be invested into it. As being reported, the design and development of the test set (Set C2) was mostly based on the available resources, specifically the graduate students taking the 519 course – Language Assessment and Testing. The test is managed within some limited financial supports. Therefore, a big question to the EPT test administers is how to allocate efforts and relevant sources among different stages of the test development in order to assure the quality of the test. Based on the findings in the discussion above, an important issue that should be focused is the revision and completion of the test specification. A new revised specification should address what aspects of language are measured in the test, and how to measure them. These decisions should then be driven by the test purpose.

Finally, within the scope of this study, only some assumptions underlying the first four warrants in the structure of the validity argument for the EPT Listening test Fall 2010 are examined. As a result, a number of assumptions for these four warrants, and the other two warrants of the last two inferences are still open for future investigations in order to provide more evidences to strengthen the validity argument for its test use and interpretation. For example, for the first inference about domain definition, a domain analysis should be carried out

in order to identify typical academic listening skills and tasks in the target domain of the EPT Listening test, which will be simulated in the test.

CHAPTER 5: CONCLUSION

This chapter contains three main contents. It first starts with a brief overview of the findings of the study leading to some implications for the operation of the EPT Listening test in specific, and the EPT test at ISU in general. Some limitations on the implementation and the results of the study will be followed. The chapter closes with the presentation of several suggestions for future research.

Overview of findings and implications of the study

The study is the first attempt to examine the reliability and validity issues of the EPT Listening Fall 2010 test in specific, and the EPT test in general. Noticeably, it tries to bring the latest validation approach in language testing which uses the concept of an interpretative argument, into practice to investigate a mid-stakes test. Accordingly, a framework of the interpretative argument for the EPT Listening test and its Fall 2010 administration is proposed with the inclusion of six inferences and corresponding warrants. The main purpose of the study is to provide backings both for and against some important assumptions underlying the first four warrants, which then become the research questions for the study.

The overall findings of the study show that the validity argument for the EPT Listening Fall 2010 test in specific, and the EPT test in general is not strong enough, and commands its test developers and administrators to make a lot of more efforts for improvements and further investigations into other assumptions of these four warrants under examination. These results also lead to some implications on how to revise the EPT Listening test, and how to use its test scores in turn.

First, the analysis results point out that the test design and development express numerous shortcomings, which need to be addressed in order to assure the reliability and validity of its measurement. Significantly, the test specification should be put as the first priority for innovation. Being found to be authentic, the test booklet itself (Set C2) is proven to require more revisions on listening texts (Lecture 4), question formats, test items and scoring methods. Moreover, the test task analyses reveal the necessity of using the test purpose to orientate all the stages, especially the very first stage in the operation of the test, which are also linked to the inferences constructing the structure of the validity argument for the test (Bachman, 2004; Chapelle et al, 2003).

Secondly, in terms of reliability, the reliability estimates of the test set (Set C2) and the score distribution of the EPT Listening Fall 2010 administration are found to be acceptable for a mid-stakes test. In other words, the test is shown to be able to discriminate the EPT Fall 2010 test takers into different groups, and the scoring of their performances by the test items in the test is seen to be reasonably consistent on the overall. However, the test still needs revising in order to attain the expected reliability index of 0.8.

Thirdly, the measurement of the EPT Listening test at ISU is found to correlate with the measurement of the TOEFL test - an internationally standardized test which assesses the same constructs of academic language skills. However, the strengths of their correlations (EPT Listening vs. TOEFL iBT Listening, EPT Listening vs. TOEFL pBT total scores, EPT Listening vs. TOEFL iBT total scores) are not as strong as theoretically-based hypotheses. Thus, in harmony with the other first two findings, the results of the correlation analyses help to confirm the fact that the EPT Listening test in general and its test set (Set C2) in specific are helpful in giving more information about the language proficiencies of the ISU admitted international students. The provision of the EPT Listening test is necessary to supplement the interpretations and use of the TOEFL scores. However, their weak strengths signify the importance of reexamining the EPT Listening test, and the reference of the students' TOEFL reported scores for better interpretations about the students' academic listening proficiencies, especially placement decisions.

What is more, the experience of adopting the structure of an interpretive argument to develop the validity argument for the EPT Listening test at ISU gives some other implications on the operation of the test. First, the creditability of a test is evaluated based on its numerous aspects, and its stages in its operation cycle, which are logically ordered and interrelated to each other. In specific, in this study, the first warrant of the EPT Listening test (set C2) shows a number of issues, which consequently cast doubts on the interpretations of the results of later investigations into other warrants in the following inferences construing the validity argument organization. This rigidly structured validity argument, thus, helps to explain why the revisions on the test specification and the test itself (Set C2) should be prioritized. Second, the assumptions explicitly stated for each warrant in each inference give good suggestions on what further investigations as well as revisions should be carried out in order to strengthen the EPT test validity argument. However, as the EPT test is a mid-stakes test which is subject to

numerous constraints, the discussion on the findings of the investigation admits the issue of taking into the available resources into consideration to choose which assumptions in which warrant should be focused first.

Limitations of the study

As being the very first investigation into the validity issue of the EPT Listening test in specific, and the EPT test at ISU in general, there exist a number of limitations on the results produced in the analyses of the study.

The first limitation is on the unavailability of relevant sources about the EPT test history and the development of the EPT Listening Fall 2010 test. First, except from the test manual, the EPT listening test specification, and its test booklet, few written records about the EPT test history as well as the design and development of the EPT Fall 2010 test (Set C2) were provided. The investigator of the validity of the test was also not the test developers, which created a new dynamics in examining the validity argument for a test. Thus, the first warrant of the validity argument for the EPT Listening Fall 2010 test would have been under a more thorough examination with the integration of its test developers' views on how each part of the test was developed.

Finally, there are some restrictions on the interpretations of the statistical analysis results of the scores by the EPT Fall 2010 test takers. The data samples for the general statistical analysis were restricted to the test-takers of the Fall 2010 administration only. Significantly, the correlation analyses were based on the test scores of the Fall 2010 test-takers who had their TOEFL scores available. In addition, the data sets vary among the four correlation analyses. Worst, little information about the sampled test-takers' characteristics was collected. Therefore, the results should be applicable to the EPT Listening Fall 2010 administration only. Moreover, for each correlation analysis, the data score set was analyzed as the whole without dividing them into different score ranges.

Suggestions for future research

As notified above, this study accounts for a partial construction of the validity argument for the EPT Listening test at ISU based on the EPT Fall 2010 administration, and thus requires more efforts to accomplish it. Based on the assumptions listed in the proposed validity argument for the EPT Listening test in Fall 2010, numerous questions are open to future examinations. Some noticeable ones can be suggested here. First, a qualitative study should be carried out in

order to look at the strategies used during test-taking, or examinees' perceptions of authenticity relative to their language use contexts and the utility of the information provided in the test results. Another important issue that should be considered is the reliability and validity of placement decisions based on the EPT Listening scores in Fall 2010. For example, some correlation analyses between the placement decisions and the students' performances on the other tests of the same measurement, or the evaluation of the instructions on the results of placement. Also, the triangulation of different evidences as well as perspectives from different participants such as academic experts, teachers, test-takers, and other test users should be taken into account in later investigations.

In addition, future studies into the EPT test at ISU may expand their investigations into other administrations as well as other test booklets in order to build up a comprehensive and proper validity argument for the EPT Listening test and the whole EPT test at ISU. To my knowledge, there have been no previous studies, which have used the interpretative argument to investigate reliability and validity issues of a mid-stakes test in the U.S. Further studies on the validity of this test type will be of great contributions to the testing field, especially to the current English language teaching and teaching situation in the U.S.

APPENDIX 1:

Specification for the English Placement Listening test at Iowa State University

Listening Test Specifications

DRAFT – 03/22/07

Specification Number: ListenSpec01

Title of Specification: Academic Listening Test

(Related Specifications, if any): Academic Reading Test, Academic Writing Test (all for the English Placement Test at ISU)

General Description (GD):

The test takers will demonstrate their ability to listen to and comprehend short speech samples and extended academic listening passages.

Specifically, learners will demonstrate the following skills and sub-skills (see Buck, *Assessing Listening*, the following skills and sub-skills still need to be fleshed out, they are currently the same as for the reading spec)

- Synthesis of information across more than one paragraph in the text
- Recognition and recovery of information in the form of specific details
- Recognition of opinions (and distinguishing them from information presented as fact)
- Recognition of inferences drawn from statements and information presented in the text
- Identification of the meaning of key vocabulary items in the text

Input for Processing (i.e.., Reading Passages):

Format of the Input:

- Channel: aural (spoken, possibly also visual, if video is used)
- Form: language (supporting visuals such as illustrations, images, and graphs/charts also possible)
- Language: English
- Length:
  for short speech samples: ~10 – 20 words
  for extended academic texts: ~600 word.
- Type: ideally intact passages where minimal editing is necessary
- Speededness: overall test is speeded; listening passage followed by a series of questions

Prompt Attributes (PA):

The reading prompt will instruct the learner to read an article (including visuals such as images, charts, graphs, and tables). The prompt will also instruct the learner to respond to a series of questions related to the reading passage.

For all questions, four options need to be provided whereby one choice represents the correct answer, one choice is plausible (not incorrect given the context), one choice is too narrow, and one choice is too broad. The correct answer needs to be marked with an asterisk (*). [Note: The specification of the different choices will need to vary depending on the item class]

Item type descriptions (refer to the attached description and examples):

1. Basic Understanding
   - Gist (implicit/explicit; abstract/concrete)
   - Detail (implicit/explicit; abstract/concrete)
2. Pragmatic Understanding
   - Stance (e.g., rhetorical purpose)
   - Function
3. Connecting information
   - Organization (rhetorical relationship)
   - Content (link, abstract info)

(Based on Item classifications for Language listening tests)

Info for item writers:

- Create 10-12 multiple choice questions with four choices for listening
- Refer to item classes (Basic Comprehension, Inferencing, Reading to learn) and to the item possibilities included in SS
- 5-6 Basic Understanding (with two inference questions)
- 2-3 Pragmatic Understanding questions
- 2-3 Connecting information questions
- Question should be of different item types
- Sample multiple choice items (see SS) ;

Response Attributes (RA):

Students will listen to the entire listening passage (and may take notes). After that, students will respond to the individual questions by selecting one (or two) choice(s) after being prompted with a written and oral question.

Sample Item (SI):

See attached

Specification Supplement (SS):

The test will be delivered as a paper and pencil test until further notice. Note: This spec is a hybrid between the Davidson/Lynch model and the specifications as outlined by Bachman and Palmer.

APPENDIX 2:

The framework for analyzing the English Placement Listening test at Iowa State University in Fall 2010 (set C2) (Taken from Buck, 2001, p. 107)

Characteristics of the setting: It consists of all the physical circumstances under which the listening takes place.

Physical characteristics: The physical conditions include all the material and equipment resources needed for a listening test (for example, the quality of recordings, background noise, recording players and loudspeakers, the quality of video or other visual aids)

Participants: Participants need to be provided with proper instructions and the best conditions (heath, or other supports) in order to have the best performance. For example, administrators have to ensure a quiet room as playing a recording, and make sure the listeners know what to do, and that they follow the instructions.

Time of task: Although the time of the test administration is regarded to be not important, in some situations certain times may be preferable, and affect the test-takers' performances.

Characteristics of the test rubric: The test rubric includes those characteristics of the test that provide structure to the test and the tasks. According to Alderson (2000), the test rubric can provide a rationale for the activity and can function in an analogous way to the listening purpose in target-language use situations. In other words, test rubric can be structured to replicate the real-world listening activities that can help test-developers design authentic tasks and materials.

Instructions: Instructions on how to do the tasks should be clear, simple, and explicit to test-takers. Also, clear examples should be given in order to prepare the test-takers with sample items before they take the test.

Test structure: This aspect of a listening test is expressed through a test specification. Accordingly, a test specification should specify the nature, the number, the content and the organization of test tasks. For example, test items in a certain test should be distributed in a difficulty hierarchical order.

Time allotment: This indicates how much time is spent for each part of a listening test which is usually determined by the sequence of texts and tasks. Usually, a listening test comprises of one recording that includes all the listening texts, instructions, questions, response pauses and so forth, and this recording will control the test.

Scoring method: A number of issues with scoring method in a listening test should be considered. First, as the aim in test development is to ensure scores are meaningful in terms of the construct defined, it is important that tasks are scorable, and the criteria for scoring are clearly determined and consistently applied. Buck also adds that writing constructed response items is easier than writing selected response items, but scoring them is much harder than scoring the others. Next, scoring criteria should be made explicit to test-takers. For example, for multiple choice tasks, a clear instruction on how to select the answer should be given so that test-takers are aware of the appropriateness of the final answer. Also, with a general understanding of relative weighting among tasks within a listening test, test-takers will have a better strategy in structuring their time and efforts to complete the test.

Characteristics of the input: The input into a listening task consists of listening texts, instructions, questions, and any materials required by the task.

Format: The format of the input comprises of a number of aspects including representativeness of spoken language, length of listening texts, channels (video or audio format), or features of other visual or written aids such as written information, questions, pictures, diagrams, ect.

Language of input: Features of the language of the input are determined based on the listening construct defined for a listening test. They include linguistic characteristics comprising of aspects of grammatical, discourse, pragmatic and sociolinguistic knowledge, as well as other audio features such as accents, stress and intonation patterns.

Topical knowledge: This aspect of a listening test refers to the content of listening texts, which is evidenced to have a great influence on test-takers' performances. Buck suggests three ways to decrease the impact of construct-irrelevant knowledge on listening comprehension: (a) use tasks that depend on knowledge that everyone has, (b) use tasks that depend on knowledge that no one has, or (c) use tasks that depend on knowledge that has been provided in the test.

Characteristics of the expected response: There are two main aspects of an expected response of interest: format of expected response and language of expected response.

Format of expected response: This aspect refers to how the response will be structured on the continuum of totally-structured to partially-structured answers. For instance, a multiple-choice format will require no efforts for test-takers to structure the answer while it requires their great amount of efforts to structure an answer in an open-question format. Other formats used in a listening test can be drawing a picture, filling a gap in a summary, or filling a diagram.

Language of expected response: This aspect involves what kind of language (first language or second language, and spoken or written) is required to answer a question in a listening test. And another issue is how to evaluate the response in terms of correctness, and appropriateness in language use, and intelligibility or clarity in meaning.

Relationship between the input and response: There are a number of aspects to look into the relationship between the input and response including reactivity, scope, and directness of relationship between the input and response.

Interactiveness: This notion is considered as an important characteristic of a test task due to its essential role in construing construct validity. In specific, it is referred as 'ways in which the test-taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task' (Bachman & Palmer, 1996, p. 25). Buck proposes two perspectives to examine interactiveness: (1) how comprehension of a listening text decides a successful completion of a test task; (2) how representative the knowledge, skills and abilities required to complete a test task are in comparison to the knowledge, skills and abilities in the construct definition.

Directness of relationship and scope: This aspect is also known as passage dependency which refers to the extent of the influence of comprehension of a listening text on successful completion of a listening task. For example, tasks may lack passage dependency because test-takers might be able to use their background knowledge or intelligence to respond. Comprehension questions can often be answered by using common sense.

Question types/formats

Buck (2008, Chapter 5) introduces two main question types used in a listening test: (1) Comprehension questions, and (2) Inference questions.

Comprehension questions are designed to measure how well the listeners have understood the content, and can be used with a variety of text-types, and to test a wide range of knowledge, skills, and abilities. Shohamy and Inbar (1991) continued to classify them into three types of questions, i.e. global questions, local questions and trivial questions. Global questions require test-takers to synthesize information or draw conclusions while local questions ask them to locate details or understand individual words. And trivial questions are to check listeners' understanding of precise but irrelevant details which are not related to the main topic. Based on their study results on the effectiveness of question types in assessing listening proficiency, trivial questions show the least information about test-takers' listening abilities, and listening questions should focus on the key information in the text, not irrelevant detail (Shohamy & Inbar, 1991, p. 37).

Inference questions are to measure listeners' inferencing ability which is considered to be at the core of language processing. As Buck's suggestion (2001, p. 147), there are two types of inferences in test tasks. The first are inferences about what the speaker means while the second are inferences about what the test-developer expects, and what the best test-taking strategy is. Likewise, some sorts of information

that can be usually addressed by inference questions, are suggested by Buck (2001, p. 148): asking for the main idea, or gist of the spoken text, or a section of the text; asking about anything which is not clearly stated, but that is clearly and deliberately indicated by the speaker; using choices of words or tone of voice – the connotations of words is a particularly rich source of inferences; asking about any pragmatic implication, or logical entailment, that follows on from what the speaker said; asking the meaning of indirect speech acts.

Regarding test question format, there are three common question formats including (1) short-answer questions, (2) multiple-choice questions, and (3) true/false questions.

APPENDIX 3:

Summary of item difficulty and item discrimination indices of 30 items in the English Placement Listening test (Set C2) at Iowa State University

Questions	Item Difficulty	Item Discrimination	Evaluation of Difficulty	Evaluation of Discrimination
Q51	0.662	0.582	rather easy	good items
Q52	0.639	0.431	rather easy	good items
Q53	0.372	-0.059	rather difficult	poor items, to be rejected or rewritten
Q54	0.772	0.733	rather easy	very good items
Q55	0.703	0.604	rather easy	very good items
Q56	0.603	0.416	rather easy	good items
Q57	0.320	-0.247	rather difficult	poor items, to be rejected or rewritten
Q58	0.507	0.286	moderately difficult	reasonably good but possibly subject to improvement
Q59	0.480	0.251	moderately difficult	reasonably good but possibly subject to improvement
Q60	0.408	0.027	moderately difficult	poor items, to be rejected or rewritten
Q61	0.392	-0.045	rather difficult	poor items, to be rejected or rewritten
Q62	0.689	0.539	rather easy	good items
Q63	0.538	0.265	moderately difficult	reasonably good but possibly subject to improvement
Q64	0.309	-0.146	rather difficult	poor items, to be rejected or rewritten
Q65	0.678	0.604	rather easy	very good items
Q66	0.678	-0.016	rather easy	poor items, to be rejected or rewritten
Q67	0.415	0.697	moderately difficult	very good items
Q68	0.426	0.229	moderately difficult	reasonably good but possibly subject to improvement
Q69	0.586	0.467	moderately difficult	good items
Q70	0.518	0.272	moderately difficult	reasonably good but possibly subject to improvement
Q71	0.433	0.207	moderately difficult	reasonably good but possibly subject to improvement

Questions	Item Difficulty	Item Discrimination	Evaluation of Difficulty	Evaluation of Discrimination
Q72	0.678	0.178	rather easy	marginal items, usually need and subject to improvement
Q73	0.678	0.539	rather easy	good items
Q74	0.699	0.741	rather easy	very good items
Q75	0.581	0.431	moderately difficult	good items
Q76	0.164	-0.729	too difficult	poor items, to be rejected or rewritten
Q77	0.507	0.135	moderately difficult	marginal items, usually need and subject to improvement
Q78	0.861	0.863	too easy	very good items
Q79	0.198	-0.564	too difficult	poor items, to be rejected or rewritten
Q80	0.401	-0.009	moderately difficult	poor items, to be rejected or rewritten

APPENDIX 4:

Results of test item analysis of 30 items in the English Placement Listening test at Iowa State University (Set C2) in terms of setting, test rubric, input, and expected response

Listening claim (General Description): The test takers will demonstrate their ability to listen and comprehend short speech samples and extended academic listening passages.				
Subclaims (Buck, 2008)	(1) Basic comprehension: - Recognition and recovery of information in the form of specific details - Identification of the meaning of key vocabulary items in the text		(2) Pragmatic understanding: - Recognition of opinions (and distinguishing them from information presented as fact)	(3) Connecting information: - Synthesis of information across more than one paragraph in the text. - Recognition of inferences drawn from statements and information presented in the text.
	(1) Setting	**(2) Test rubric**	**(3) Input**	**(4) Expected response**
Physical characteristics	the quality of video and recordings are good	**Instructions:** instructions are well prepared, including an an introduction about the test with its components, and instructions on how to do each test. Instructions are given in both spoken and written forms, and provided with examples.	**Format:** All the four listening texts have a lead-in by a narrator and use video, or visual aids besides the audio channel. **Representativeness of spoken language:** All the four lectures used are authentic, and the professors represent two typical accents of native English speakers (North American English, and British English). Lectures 1 and 3 are given by two professors at Standford University, lecture 2 by a professor at ISU, and lecture 4 by a professor from University College in London. **Length:** Lecture 1 (537 words in 2:58 minutes). Lecture 2 (358 words in 2:33 minutes). Lecture 3 (478 words in 3:28 minutes). Lecture 4 (1299 words in 7 minutes) **Channels:** All the first three lectures start with a slide including the name of the professor, and the topic of the lecture. All the four lectures use video, captions, and images along the listening text. However, the captions appear to be quite small on the screen for the back-rows in the testing room.	**Format of expected response:** multiple-choice (partially structured answers), the scoring rubric is dichotomous (Right/Wrong; 1-0) **Language of expected response:** all the questions and given choices are in English.

Participants	some late comers might still suffer from jet lag (daytime and evening time for testing)	**Test structure:** The test specification is claimed to be based on the framework of academic listening given by Buck (2001), and the hybrid of the test-task characteristics (Bachman, & Palmer, 1996) and Davidson and Lyn's model. The specification includes: (1) the nature of the test, (2) the number of test-tasks, including the number of questions, and the distribution among different question types, (3) the organization of the test. Some characteristics of the test task, and the design of questions are also described. However, the test specification includes two types of listening including short listening samples, and academic lectures. However, the real test focuses only academic lecture listening.	**Language of input: Linguistic features (grammatical, discourse, sociolinguistic/pragmatic):** Each lecture involves a different way of delivery. Lectures 1, 2, and 4 are monologic while Lecture 3 is more interactive as a news report. Therefore, a number of different types of discourses are included in the listening test. **Audio features (accent, intonation, stress):** all the listening lectures are authentic, and the speakers are all native-speakers of English. Especially, lecture 2 is given by a professor working in a center in Iowa, which is expected to strongly represent the Mid-western accent.	

	other conditions are good (supported by test examiners	**Time allotment:** The test is speeded, and controlled by one recording including instructions, listening texts, questions, and response pauses. In other words, each listening passage is followed by a series of questions which are in both spoken and written forms.	**Topical knowledge**: The listening test comprises of four listening texts rendering four different academic areas including social sciences (lecture 1, lecture 4), natural science and engineering (lecture 2, and lecture 3). Lecture 1: Team composition Lecture 2: Research in plant pathology Lecture 3: Technology and Engineering (Car driving simulation) Lecture 4: How the internet enables intimacy	
Time of the test administration	It might be disadvanta geous for some new arrivals	**Scoring method:** The test uses the multiple-choice format, and distributes the computer form for test-takers to record their answers. There is no weighting among questions. The test-takers are also not informed of the scoring method, and the possible score for each part (which is not important due to no weighting)		

APPENDIX 5:

Results of test item analysis of 30 items in the English Placement Listening test at Iowa State University (Set C2) in terms of the relationship between the input and response, question types and formats

(1) Question types
(2) Interactiveness
(+) To make the correct choice requires a comprehension of the listening text, and highly representative academic knowledge, skills, and abilities.
(-) To make the correct choice does not require a comprehension of the listening text, and highly representative academic knowledge, skills, and abilities.
(3) Directness of relationship and scope
(+) The successful completion of a listening task is dependent on the comprehension of a listening text.
(-) The successful completion of a listening task is not dependent on the comprehension of a listening text, but background knowledge or intelligence.

Results of test item analysis of 30 items in the English Placement Listening test at Iowa State University (Set C2)

Lectures	**Lecture 1 (Q.51-56)**			**Lecture 2 (Q.57-62)**			**Lecture 3 (Q.63-70)**			**Lecture 4 (Q.71-80)**		
	Question types	Relationship between the input and response		Question types	Relationship between the input and response		Question types	Relationship between the input and response		Question types	Relationship between the input and response	
Question No	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	Inference (main idea)	+	+	Comprehension (global)	(+)	+	Comprehension (local)	+	+	Inference (main idea)	+	+.
2	Comprehension (global)	+	+.	Inference	+	+	Inference	+	+	Comprehension (Global)	+	+.
3	Comprehension (global)	+	+	Inference	+	+	Inference	+	-	Comprehension (Global)	+.	+
4	Inference	+	+	Comprehension (local)	+	+	Inference	+.	-.	Inference	+	+.
5	Inference	+	+.	Inference	+	+	Inference	+	+	Inference	+	+
6	Comprehension (global)	+	+	Comprehension (global)	+	-	Comprehension (local)	+	+	Comprehension (global)	+	+.
7				Notice: Q.59 and 61 rely on the similar piece of information in the listening text.			Inference	+	+	Inference	+	+.
8							Comprehension (global)	+	+	Comprehension (global)	+	+
9										Comprehension (trivial)	-	+
10										Comprehension (trivial)	-	+

APPENDIX 6:

Summary of the comparison in the test format between TOEFL pBT and TOEFL iBT

Part	TOEFL pBT	TOEFL iBT
Listening	50 questions	33-34 questions
	Three types of questions are presented in three separate parts. Part A has short conversations; Part B has long conversations and class discussion; Part C has mini-talks and lectures.	Two types of questions are presented in six sets: The first sets each have a long conversation. The next sets each have one lecture.
	The talks and lectures are about 2 minutes long.	The lectures are about 5 minutes long.
	There are no pictures of visual cues.	Each conversation and lecture begins with a picture to provide orientation. There are several pictures and visual cues with lectures.
	You hear the questions, but they are not written out for you to read.	The questions are written out on the computer screen for you to read while you hear them.
	Everyone taking the TOEFL proceeds at the same pace. You cannot pause the tape.	You may control the pace by choosing when to begin the next conversation or lecture.
	The section is timed. At the end of the tape, you must have completed the section.	The section is timed. A clock on the screen shows the time remaining for you to complete the section.
	You may not replay any of the conversations or lectures.	You may not replay any of the conversations or lectures.
	All of the questions are multiple-choice.	Most of the questions are multiple-choice, but some of the questions have special directions.
	Every question has only one answer. You answer on a paper Answer Sheet, filling in ovals marked A,B,C and D.	Some of the questions have two or more answers. You click on the screen in the oval that corresponds to the answer you have chosen, or you follow the directions on the screen.
	You can return to previous questions, erase, and change answers on your answer sheet.	You cannot return to previous questions. You can change your answer sheet before you click on OK. After you click on OK, you cannot go back.
	You may not take notes.	You may take notes while you listen to the conversations and lectures
Speaking	No questions	6 questions
		Three types of questions are presented in six sets. The first two sets have a general question; other sets have questions about campus and academic topics.
		After you see and hear the general questions, you will have 15 seconds to prepare your answers and 45 seconds to record them.

Part	TOEFL pBT	TOEFL iBT
		After you hear the campus and academic questions, you will have 20-30 seconds to prepare each answer and 60 seconds to record it.
Structure	40 questions	
	Two types of questions are presented in separate parts. Part A has incomplete sentences, and Part B has sentences with underlined words and phrases.	There is NO structure section.
	All of the questions are multiple-choice.	
	Everyone taking the TOEFL answers the same questions	
	Every question has only one answer. You answer on a paper Answer Sheet, filling in ovals marked A,B,C and D.	
	You have 25 minutes to complete the section	
	You can return to previous questions, erase, and change answers on your answer sheet.	
	The score on the Structure section is not combined with the score on the essay in the Test of Written English (TWE)	
Reading	50 questions	36-39 questions
	There are five reading passages with an average of ten questions after each passage.	There are three reading passages with an average of 12-13 questions after each passage.
	The passages are about 250-300 words in length.	The passages are about 700-800 words in length.
	Everyone taking the TOEFL answers the same questions	You will answer the same questions as others who take the same form of the test.
	There are no pictures of visual cues.	There may be pictures in the text and questions that refer to the content of the reading passage.
	All of the questions are multiple-choice.	Most of the questions are multiple-choice, but some of the questions have special directions.
	Every question has only one answer. You answer on a paper Answer Sheet, filling in ovals marked A,B,C and D.	Some of the questions have two or more answers. You click on the screen in the oval that corresponds to the answer you have chosen, or you follow the directions on the screen.
	You can return to previous questions, erase, and change answers on your answer sheet.	You can return to previous questions, change answers and answer questions you have left blank, but you cannot return to passages in a previous part.
	There is no glossary	There may be a glossary of technical terms.
	You may not take notes.	You may take notes while you read.

Part	TOEFL pBT	TOEFL iBT
	1 question	2 questions
	The essay, also called the Test of Written English (TWE), is offered five times each year. The test-taker must select a TOEFL test date when the TWE is scheduled of he or she needs an essay score.	The writing section is required.
	There is only one topic for each essay.	There are two topics. The first is independent writing while the second one is based on both a lecture and a reading passage.
	Everyone takng the TOEFL writes an essay about the same topic.	Everyone taking the same form of the TOEFL will write about the same topics.
	The test-takers do not know any of the topics for the essay before the test administration.	At this point, no writing topics have been published; however, the essay topics previously published for the computer-based TOEFL are good practice for the general topic essay.
	Most of the topics ask the test-taker to agree or disagree with a statement or to express an opinion.	The topic for the independent writing task asks you to agree or disagree with a statement or to express an opinion. The integrated task refers to topics from a lecture and a reading passage.
	The topics are very general and do not require any specialized knowledge of the subject to answer them.	The independent topics are very general and do not require any specialized knowledge of the subject to answer them. Technical words are explained in the text or in a glossary for the integrated topics.
	30 minutes to complete the essay	The test-takers have 30 minutes to complete the independent writing task, and 20 minutes to complete the integrated writing task.
	The test-takers handwrite their essays on paper provided in the test materials.	The test-takers should type the writing samples on the computer.
	The test-takers have one page to organize their essays. The page is not graded.	The test-takers have paper to take notes and organize the writing. The notes and outlines are not graded.
	The essay will not be scored for neatness, but the readers must be able to understand what have been written.	The essay will not be scored for neatness, but the readers must be able to understand what have been written.
	The essay is holistically scored on the scale of 1 to 5.	A scale from 0 to 5 is used to grade writing samples.
Writing	The score is reported separately from the TOEFL score. It is not included in the computation of the total TOEFL score and does not affect the score on the multiple-choice TOEFL.	The score is reported as a separate writing section score.

REFERENCES CITED

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (eds.) *A new decade of language teaching research.* Alexandria, VA: TESOL, 46-57.

American Psychological Association (APA). (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Bachman, L.F. (2004). *Statistical analyses for language assessment.* Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language assessment quarterly*, 2 (1), 1-34.

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

Bachman, L. F., Kunnan, A., Vanniaranjan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency tests. *Language testing*, 5, 128-159.

Brown, G. (1990). *Listening to spoken English* (2nd ed.). London: Longman.

Brown, J. D. (1989). Improving ESL Placement tests using two perspectives. *TESOL Quarterly*, 23(1), 65-83.

Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (eds.). *Interfaces between second language acquisition and language testing research,* 32- 70. Cambridge: Cambridge University Press.

Chapelle, C. (1999). Validity in language assessment. *Annual review of applied linguistics,* 19, 254-272.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. New York : Cambridge University Press.

Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle, J. Jamieson, & M. Enright (eds.), *Building a validity argument for the test of English as a foreign language,* 319-352. London: Routledge.

Chapelle, C., Jamieson M., & Enright, K. (eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language* . New York: Routledge.

Chapelle, C., Enright, K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational measurement: Issues and practice,* 20, 1, 3-13.

Chapelle, C., Chung, Y, Hegelheimer, V., Pendar, N. & Xu, J. (2010). Towards a computer-delivered test of productive grammar ability. *Language Testing,* 20, 1-27.

Chiang, C., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly,* 26(2), 345-374.

Cronbach, L. (1988). Five perspectives on validity argument. *Test validity,* 3-17.

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

Crooks, T. J., Kane, M. T. & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in education: Principles, policy & practice*, *3*(3), 265-286.

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

Douglas, D. (ed.). (2003). *English language testing in U.S colleges and universities* (2nd ed.). Washington, D.C: Association of International Educators.

Douglas, D. (2009). *Understanding Language Assessment.* London: Hodder.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.*).* New Jersey: Prentice-Hall.

Educational Testing Service (ETS). (2011). About the test. Retrieved on January 13th, 2011 from http://www.ets.org/toefl/ibt/about

Educational Testing Service (ETS). (2011). About the test. Retrieved on January 13th, 2011 from http://www.ets.org/toefl/pbt/about

Feyten, C. (1991). The power of listening ability: an overlooked dimension in language acquisition. *The modern language journal,* 75 (2), 173-180.

Flowerdew, J. (1994). *Academic listening: research perspectives.* Cambridge: Cambridge University Press.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language testing,* 14(2), 113-139.

Goodbody, M.W. (1993). Letting the students choose: a placement procedure for a pre-sessional course. In Blue, G.M. (eds), *Language, learning and success: Studying through English.* London: Modern English Publications and the British Council, 49-57.

Hanson, C. & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (ed.), *Academic listening: research perspectives.* Cambridge: Cambridge University Press.

Jensen, C., & Hanson, C. (1995). The effect of prior knowledge on EAP listening performance. *Language testing,* 12 (1), 99-119.

Kane, M. (1992). An argument-based approach to validity. *Psychological bulletin,* 112(3), 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of educational measurement,* 38(4): 319-342.

Kane, M. (2002). Validating high stakes testing programs. *Educational measurement: Issues and practice,* 21(1), 31-41.

Kane, M. (2004). The analysis of interpretive arguments: some observations inspired by the comments. *Measurement: Interdisciplinary research and perspective,* 2(3), 192-200.

Kane, M. (2006). Validation. In R. Brennon (ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger, 17-64.

Kane, M., Crooks, T., et al. (1999). Validating measures of performance. *Educational measurement: Issues and practice,* 18(2), 5-17.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL*, 29(2), 135–149.

Lado, R. (1961). *Language testing : The Construction and use of foreign language tests : A teacher's book*. Inglaterra : Longmans, Green and Company.

Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test. *Journal of mixed methods research ,* 1(4), 366-389.

Messick, S. (1988). Validity. In R. L. Linn (eds.), *Educational measurement* (3rd ed.). New York: Macmillan, 13-104.

Messick, S. (1989). *Validity.* Macmillan: American Council on Education.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice,* 14 (4), 5-8.

Mislevy, R. L., Steinberg, et al. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research & perspective,* 1(1), 3-62.

Mislevy, R. L. (2003). *Argument substance and argument structure in educational assessment.* CSE Technical Report 605. Los Angeles: Center for the study of evaluation.

Richards, J. C. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly,* 17 (2), 219-240.

Rost, M. (2002). *Teaching and researching listening.* New York: Pearson Education.

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of educational measurement,* 33(3), 271-290.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (ed.). *Review of research in education,* 19, 405-450. Washington, D.C: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational measurement: Issues and practice,* 16 (2), 5-8.

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing: A resource for teachers and program administrators.* Alexandria, VA: TESOL Publications.

Toulmin, S., Richard, R., & Allan, J. (1979). *An introduction to reasoning* (2nd ed.). New York: Macmillan.

Toulmin, S. (2003). *The uses of arguments* (eds.). Cambridge: Cambridge University Press.

Truman, W. L. (1992). College placement testing. *AMATYC Review,* 13, 58-64.

Usaha, S. (2000). Effectiveness of Suranaree University's English placement test. Suranaree University of Technology. Retrieved on 12th September 2010 from http://hdl.handle.net/123456789/2213

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

Wesche, M., Paribakht, T. S., & Ready, D. (1993). A comparative study of four ESL placement instruments. In Micheal Milanovic & Nick Saville (eds), *Studies in language testing 3: performance testing, cognition, and assessmen*t. Cambridge: Cambridge University Press.

Wesche, M. (1987). Second language performance testing: The Ontario test of ESL as an example. *Language testing,* 4, 28-47.