

A COMPARABILITY AND CLASSIFICATION ANALYSIS OF
COMPUTERIZED ADAPTIVE AND CONVENTIONAL PAPER-BASED
VERSIONS OF AN ENGLISH LANGUAGE PROFICIENCY READING
SUBTEST

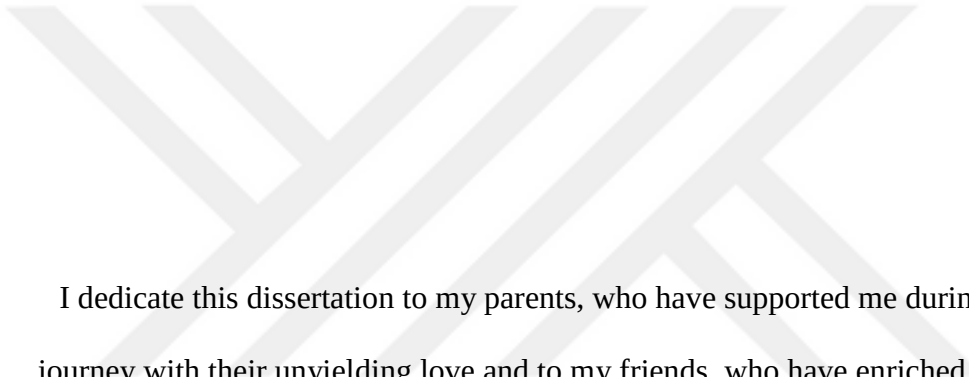
A DOCTORAL DISSERTATION

BY

ELİF KAYA

THE PROGRAM OF CURRICULUM AND INSTRUCTION
İHSAN DOĞRAMACI BILKENT UNIVERSITY
ANKARA

JANUARY 2022



I dedicate this dissertation to my parents, who have supported me during this long journey with their unyielding love and to my friends, who have enriched my soul and inspired me to complete this research.



A Comparability and Classification Analysis of Computerized Adaptive and
Conventional Paper-Based Versions of an English Language Proficiency Reading
Subtest

A Dissertation

Submitted in Partial Fulfilment of the Requirements

for the Degree of Doctor of Philosophy

in

Curriculum and Instruction

by

Elif Kaya

The Graduate School of Education

of

İhsan Doğramacı Bilkent University

Ankara

January, 2022

İHSAN DOĞRAMACI BILKENT UNIVERSITY

GRADUATE SCHOOL OF EDUCATION

A Comparability and Classification Analysis of Computerized Adaptive and Conventional Paper-Based Versions of an English Language Proficiency Reading Subtest

Elif Kaya

December 2021

I certify that I have read this doctoral dissertation and have found that it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy in Curriculum and Instruction.

Asst. Prof. Dr. İlker Kalender (Advisor)

I certify that I have read this doctoral dissertation and have found that it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy in Curriculum and Instruction.

Asst. Prof. Dr. Aikaterini Michou (Examining Committee Member)

I certify that I have read this doctoral dissertation and have found that it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy in Curriculum and Instruction.

Asst. Prof. Dr. Jennie Farber Lane, (Examining Committee Member)

I certify that I have read this doctoral dissertation and have found that it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy in Curriculum and Instruction.

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

I certify that I have read this doctoral dissertation and have found that it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy in Curriculum and Instruction.

Assoc. Prof. Dr. Semirhan Gökçe, Niğde Ömer Halisdemir University (Examining Committee Member)

Approval of the Graduate School of Education

Prof. Dr. Orhan Arıkan (Director)

ABSTRACT

A COMPARABILITY AND CLASSIFICATION ANALYSIS OF COMPUTERIZED ADAPTIVE AND CONVENTIONAL PAPER-BASED VERSIONS OF AN ENGLISH LANGUAGE PROFICIENCY READING SUBTEST

Elif Kaya

Ph.D. in Curriculum and Instruction

Supervisor: Asst. Prof. Dr. İlker Kalender

January 2022

The current study compares the computerized adaptive test (CAT) and paper-based test (PBT) versions of an English language proficiency reading subtest in terms of psychometric qualities. The study also investigates classification performance of CATs not designed for classification purposes with reference to its PBT version. Real data-based simulations were conducted under varying test conditions. The results demonstrate that ability levels estimated by CATs and PBT are similar. A relatively larger item reduction can be obtained with 0.50 and 0.40 standard error thresholds and CATs terminated with 20, 25, and 30 items performed well with acceptable SE values. Reliability of CAT ability estimates was comparable and highly correlated with PBT estimates. For classification analysis, classification accuracy (CA) and classification consistency (CC) was also estimated using the Rudner method. Classification analyses were conducted on single and multiple cut-off points. The results showed that the use of a single cut-off score produced better classification performance, particularly for high and low ability groups. On the other hand, the use of multiple cut-off scores simultaneously yielded significantly lower classification performance. Overall, the results highlight the potential for CATs not designed specifically for classification to serve classification purposes and indicate avenues for further research.

Keywords: Computerized adaptive testing, comparability, classification consistency, classification accuracy, Rudner approach, language proficiency

ÖZET

BİR İNGİLİZCE YETERLİLİK SINAVI OKUMA ALT TESTİNİN BİLGİSAYAR ORTAMINDA BİREYSELLEŞTİRİLMİŞ TEST VE GELENEKSEL KAĞIT KALEM TEST UYGULAMALARININ KARŞILAŞTIRILABİLİRLİĞİ VE SINIFLANDIRMA ANALİZİ

Elif Kaya

Doktora, Eğitim Programları ve Öğretim

Tez Yöneticisi: Dr. Öğr. Üyesi İlker Kalender

Ocak 2022

Bu araştırma bir İngilizce yeterlilik okuma alt testinin bilgisayar ortamında bireyselleştirilmiş test (CAT) ve geleneksel kağıt kalem test olarak uygulanmasından elde edilen puanları çeşitli psikometrik özellikler açısından karşılaştırmakta ve sınıflandırma yapmak amacıyla tasarlanmamış tipik CAT uygulamalarının sınıflandırma performansını kağıt kalem testiyle kıyaslayarak incelemektedir. Bu amaçla farklı test sonlandırma kuralları uygulayarak gerçek veriye dayalı simülasyonlar yapılmıştır. Sonuçlar CAT ve PBT uygulamalarından elde edilen yetenek kestirimlerinin benzer olduğunu ortaya koymuştur. CAT uygulamasında 0,50 ve 0,40 standart hata eşikleriyle sonlandırılan testlerde soru sayısında göreceli olarak daha fazla bir azalma elde edildiği ve sabit soru sayısı kuralıyla sonlandırılan testlerde 20, 25 ve 30 soru sayılarının kabul edilebilir standart hata değerleri ile birlikte iyi sonuçlar verdiği gözlemlenmiştir. CAT uygulamalarından elde edilen yetenek kestirimlerinin güvenilirliği ile kağıt kalem test uygulamasından elde edilen kestirimlerin güvenilirliği benzer ve yüksek düzeyde ilişkili çıkmıştır. Sınıflandırma analizi kapsamında, çalışmada sınıflandırma doğruluğu (CA) ve tutarlılığı (CC) Rudner yöntemi ile kestirilmiştir. Farklı sayı ve noktalarda kesme puanı göz önünde bulundurulmuştur. Sonuçlar tek bir kesme puanının kullanılmasının, özellikle yüksek ve düşük yetenek gruplarında, daha iyi bir sınıflandırma performansına yol açtığını göstermiştir. Öte yandan, tek seferde birden çok kesme puanının aynı anda kullanılmasının ise sınıflandırma performansını önemli derecede düşürdüğü görülmüştür. Genel olarak, sonuçlar sınıflandırma yapmak için tasarlanmamış tipik CAT uygulamalarının da sınıflandırma potansiyelini ortaya koymuş ve ileride yapılabilecek olası araştırma alanlarını da belirlemiştir.

Anahtar kelimeler: Bilgisayar ortamında bireyselleştirilmiş testler, sınıflandırma tutarlılığı, sınıflandırma doğruluğu, Rudner yaklaşımı, dil yeterliliği

ACKNOWLEDGEMENTS

I would like to express my sincerest thanks to Assist. Prof. Dr. İlker Kalender for being my supervisor and for his support, guidance and patience throughout the challenging process of building this dissertation. I would also like to thank to the members of my dissertation committee, Assist. Prof. Dr. Aikaterini Michou and Prof. Dr. Giray Berberoğlu for many insightful discussions and sharing their expertise. I would like to thank Dr. Stefan O’Grady for his invaluable support and feedback at the publication process of our article. His expert guidance was tremendously helpful throughout the process of writing our co-authored article in *Language Testing*. I am grateful to you all for your time and effort. Last but not least, I wish to thank my family, who has always supported and given me strength to continue during this long journey.

TABLE OF CONTENTS

ABSTRACT.....	iii
ÖZET.....	iiiv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES	iiix
LIST OF FIGURES	ix
CHAPTER 1: INTRODUCTION	1
Background	1
Measurement and Assessment in Language Proficiency.....	2
Computerized Adaptive Test (CAT).....	4
Comparability of Scores across Different Test Delivery Modes.....	9
Classification with CAT	10
Problem	13
Purpose	15
Research Questions	16
Significance	17
CHAPTER 2: REVIEW OF RELATED LITERATURE	19
Introduction	19
Classical Test Theory (CTT).....	19
Item Response Theory (IRT).....	23
Item and Test Information	28
Ability Estimation in IRT	32
Assumptions of IRT.....	34
Limitations of IRT	38

Computerized Adaptive Testing.....	39
Ability Estimation.....	43
Item Selection	44
Test Termination.....	46
Language Proficiency	47
Language Testing	50
Classification of Examinees	56
Summary.....	61
CHAPTER 3: METHOD	63
Introduction	63
Research Design	63
Context	65
Participants	66
Instrumentation.....	67
Method of Data Collection	69
Methods of Data Analysis	70
Preliminary Analysis.....	70
Fit of Model-Data Analysis	72
Post-Hoc Simulations	73
Comparability Analysis	76
Classification	77
CHAPTER 4: RESULTS	79
Introduction	79
Comparability Analysis	79
Classification Performance of CAT	87

Percentile Rank Analysis in CAT Results with Reference to PBT	87
Estimating Classification Accuracy (CA) and Classification Consistency (CC)	92
Summary	100
Comparability	100
Classification	100
CHAPTER 5: DISCUSSION	103
Introduction	103
Overview of the Study	103
Discussion of Major Findings	105
Implications for Practice	114
Implications for Further Research	116
Limitations	119
Conclusion	118
REFERENCES	122
APPENDIX A. Correctly Classified Examinees into the Same Decile by SE-based CATs	150
APPENDIX B. Correctly Classified Examinees into the Same Decile by FL CATs	151
APPENDIX C. Details of the Twenty-Five Most Frequently Used Items in CAT Simulations	152
APPENDIX D. Details of the Items in PBT	154

LIST OF TABLES

Table		Page
1	Summary of IRT-Based Approaches to Estimating CA/CC...	61
2	Subtests of English Language Proficiency Test.....	69
3	Descriptives of Ability Estimates for Full-length PBT.....	72
4	Descriptives of Ability Estimates by CATs Averaged over 100 Replications.....	80
5	Means (SDs) of the Ability Estimates by CATs Averaged over 100 Replications.....	82
6	Descriptives of the Number of Items Used in SE CATs	82
7	Descriptives of Examinees Administered 35 items in SE CATs.....	83
8	Descriptives of Ability Estimations at ≤ 0.40 SE for FL CATs.....	84
9	Descriptives of Ability Estimations at ≤ 0.50 SE for FL CATs.....	85
10	Mean SE Values for FL CATs.....	86
11	Correlations between PBT and CAT Ability Estimations.....	86
12	Correlations between CAT Ability Estimations and Raw Scores of Listening Subtest.....	87

LIST OF FIGURES

Figure		Page
1	Processes of CAT.....	6
2	Two Item Characteristic Curves for 2-PL Model.....	31
3	Test Information Function.....	32
4	Schematic Representation of a Computerized Adaptive Test..	42
5	Illustration of CA and CC under Rudner's Method	60
6	Illustration of CA/CC with Lee Method.....	61
7	Histogram of Ability Estimation by Full Length PBT.....	72
8	Item Information and Posterior SD Curves for PBT.....	74
9	Histogram of SE Values for PBT Ability Estimations.....	85
10	Percentages of Correctly Classified Examinees into the Decile by SE-based CATs.....	89
11	Percentage of Examinees Classified into One Ability Group Above or Below by SE-based CATs.....	90
12	Percentage of Examinees Classified into One Ability Group Above/Below by FL CATs.....	91
13	Percentage of Examinees Classified into One Ability Group Above or Below by FL CATs.....	92
14a-d	CA/CC Values with 9 Cut-Off Points.....	95
15a-d	CA/CC Values with 5 Cut-off Points.....	98
16	CA/CC Values for Simultaneous Classifications with 5 and 9 Cut-Off Points.....	100

CHAPTER 1: INTRODUCTION

Background

Measurement plays a crucial role in the sciences, engineering fields as well as many everyday activities. Linn and Gronlund (2000) define psychological measurement as the process of obtaining a numerical description of the level to which an individual has a specific trait. In other words, it involves assigning numbers to a particular type of measurement with reference to a specific rule such as awarding points for various aspects of an essay. In education, traits like aptitude, opinion, scholastics ability, attitude or achievement can be measured using different instruments like questionnaires. These instruments have a multitude of purposes such as description, prediction, assessing individual differences, objective evaluation, domain estimation, mastery decisions and diagnosis (Wiersma & Jurs, 1990).

The term assessment, on the other hand, is far more comprehensive. It involves various procedures to collect information about student performance (Linn & Gronlund, 2000). These procedures range from traditional paper-based tests to extended response essays and performances of authentic tasks such as laboratory experiments. The ultimate purpose of assessment is gathering, interpreting, and evaluating evidence of learning, which allows educators to discover student needs, provides feedback about how effective a particular teaching practice is, guides curriculum design and is a means of evaluation and accountability for the system (Gitsaki & Coombe, 2016). Compared to measurement, assessment may include both quantitative and qualitative descriptions of students as well as value judgments regarding the results. Assessment is not necessarily based on measurement, but if it

is, it goes beyond mere quantitative descriptions (Linn & Gronlund, 2000). Finally, it is worth noting that both measurement and assessment are generally designed around and based on the objectives specified in the curriculum of a specific educational program. As well as indicating the students' progress in terms of the target knowledge and skills in the curriculum, measurement and assessment also show how well a particular curriculum is performing in supporting the students' acquisition/mastery of the desired knowledge and skills in the students.

Another common use of educational measurement and assessment is to classify examinees. Classification is regarded as high-stakes because errors in classification may cause wrong decisions (Eckes, 2017) and deprive individuals of important opportunities such as educational or career development (Zhang, 2010). This dissertation investigates the classification performance of computerized adaptive tests (CAT) in language assessment in higher education following a comparability analysis across two types of testing: paper-based test (PBT) and CAT.

Measurement and Assessment in Language Proficiency

Language proficiency tests measure the extent to which an individual is able to use language to communicate. Language assessment used for initial screening, placement, or progression in a language course plays a crucial role in language learners' educational experience and learning outcomes. Assessment is one of the most demanding tasks for language educators to complete as designing valid and reliable assessment tools requires specialized skills. In addition, decisions made based on assessment, particularly on high-stakes exams, can have a lasting effect on students' life.

Modern instruments designed to measure mastery of a second Language such as English involve students from various language learning backgrounds and is often

used to determine the minimum proficiency level these students need to study in English-medium institutions (EMI). These language proficiency exams are administered many times during an academic year and are referred to as high-stakes tests as they are critically important for important enrollment and graduation decisions. IELTS, paper-based TOEFL and TOEFL IBT can be listed as the most commonly used test of English. These exams are taken by many people all around the world for various purposes such as studying abroad. In Turkish higher education, with the growing number of EMI universities in this context, language proficiency exams have gained more importance and it has led to increased scrutiny of the development of in-house English language proficiency tests (Selvi, 2014).

Conventional PBTs have been commonly used in language testing. Such tests are often based on Classical Test Theory (CTT). Despite the usefulness of CTT in testing, researchers and assessors have found several limitations with this method of testing (Eleje et al., 2018). One of the major limitations of conventional tests constructed according to CTT is that while they provide insights into the mean responses of the participants, they yield poor measurements for individuals whose ability levels deviate from that mean, especially when items are not well-calibrated. (Weiss, 2004). However, in practice, the range of abilities to be measured is wide and covering the whole range is an important requirement for measurement precision. Another limitation of conventional tests is that ability and item parameters in CTT depend on total score and sample (Cronbach, 1990; Hambleton et al., 1991). In other words, item difficulty and item discrimination as classical item statistics are based on the examinee samples from which they were obtained from (Cappelleri et al., 2014). An inevitable result of this is that item statistics are useful if tests are constructed for an examinee sample that has similar characteristics to the one used in

item calibration. However, it is not always easy to be certain that the sample of examinees for whom a test is targeted is similar to the one used in obtaining the item statistics.

Apart from these limitations, there are other issues associated with conventional PBTs. Common ones can be listed as high number of examinees, the need for high numbers of questions to cover various proficiency levels, undermining motivation in students that are both high achieving and low performing as they have to answer questions which are far below or way beyond their proficiency level. Other issues are the logistics of administering a high-stakes proficiency exam with the required number of exam rooms, invigilators, security cameras, exam papers, optic forms, envelopes, pens and pencils, the demand to administer the test in a standardized manner, which makes the process time-consuming and costly (Christenson et al., 2007; Weiss, 2004).

Given that one form of assessment is never enough to give all the information needed to measure students' progress and language proficiency levels, teachers, test developers, administrators, and policy makers need to be aware of approaches and methods that have the potential to generate valid and reliable evidence of student learning and integrate various assessment techniques into their practice. The following section discusses an alternative to PBT that is being utilized by an increasing number of test centers, yet still has the need for research to ensure reliability and validity.

Computerized Adaptive Test (CAT)

CAT is regarded as a strong alternative to PBT and it offers solutions to most of the limitations encountered when PBT is used. Using algorithms, CAT tailors the test for examinees by selecting items based on examinees' responses to previous

items (Dunkel, 1999). A typical CAT is based on Item Response Theory (IRT), also known as modern test theory. One of the benefits of IRT is that item statistics do not vary depending on examinee samples (Eleje et al., 2018). IRT relates ability level of examinees to item parameters (i.e., difficulty, discrimination, and pseudo-guessing). IRT makes attempts to match several item parameters (i.e., item difficulty, item discrimination and item pseudo-guessing) with the probability of giving a correct answer to that item (Cheng & Morgan, 2012). All CATs are designed to use some or all of these parameters. A large pool of items is needed by IRT to match test items to ability levels, which makes it possible to select appropriate items to measure the ability level of the examinees.

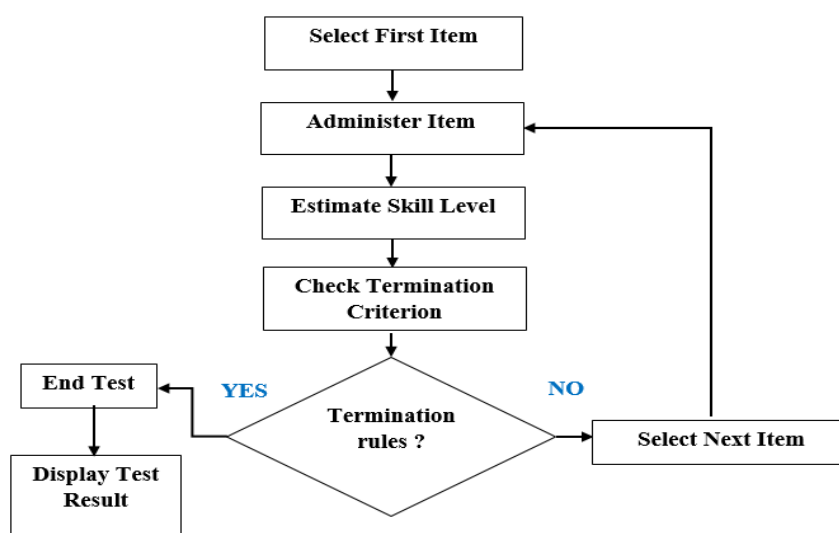
The idea of tailoring tests to the individual was first proposed by Alfred Binet with Binet IQ test (Binet & Simon, 1905). The test involved sets of test items to assess the IQ of children aged 3 years through 11 years which constituted Binet's item "bank" for his adaptive test. Items were identified as suitable for that age level if around half of the children at that age level supplied an accurate answer to that. Binet's test involves a fully adaptive procedure with all its components including an item selection rule and predefined termination rule. Since Binet, computer programmes have been developed to successfully predict how the examinee would respond to test items not administered yet based on the responses given earlier, selecting the next test item by making effective use of this knowledge and assigning a numerical score at the end representing the ability of the examinee.

CAT was conceptually developed by Lord (1970). Lord also named it "tailored testing" because the test is adapted to each examinee level. It is possible to obtain relatively more precise information from CATs for the trait measured compared to conventional tests (Young et al., 1996). A typical CAT continues until it

can obtain an estimate of the proficiency of the examinee based on a predetermined level of accuracy (Wainer et al., 2000). As shown in Figure 1 (Boussakuk et al., 2020), CAT administration involves three major processes which are estimating ability estimate, checking the progress of the test against a test termination rule, and selecting an item depending on the estimated ability. These processes iterate every time an item is administered, as shown in Figure 1.

Figure 1

Processes of CAT



The primary advantage of CAT is shortening tests without a decrease in measurement precision. Another important advantage of CAT is that test scoring and ability estimation is conducted while the test is being administered. Rather than administering the same items to all examinees, CAT administers an optimal set of items tailored to each individual thereby increasing measurement efficiency (Morris et al., 2020). In addition to this, according to Hambleton et al. (1991), some of the advantages that CATs offers are as follows (Verschoor & Straetmans, 2010; Wainer et al., 2000):

- Higher test security

- Testing on demand
- No need for answer sheets or optic forms
- Test pace depends on the individual instead of standard testing time
- Immediate test scoring and reporting
- Minimum test frustration for some examinees
- Greater test standardization
- Easy removal of defective items from the item bank when identified
- More flexibility in the choice of item formats
- Less test supervision time

In contrast to important advantages, CAT also involves challenges relating to careful processes of item pool development and checking to make sure that the items are as seamless as possible. In general, a CAT that is half as long may attain the same level of measurement; therefore, each item is critical. In other words, “if an item is flawed, its impact on the estimate of the examinee’s proficiency is doubled” (Wainer et al., 2000, p. 13).

In a CAT framework, estimation of individuals’ ability is an important component. Ability estimation can be done with item parameters or pre-estimated item parameters. Four ability estimation methods are commonly used to estimate ability level of individuals: Maximum Likelihood Estimation (MLE) and three Bayesian approaches (Owen's method, Expected a Posteriori (EAP), and Maximum a Posteriori) (Wang & Vispoel, 2005). The results obtained with MLE and Bayesian estimation methods are sometimes nearly identical but these methods are conceptually different. While MLE views the parameters as quantities whose values are fixed but unknown, Bayesian methods consider the parameters as random variables with some known a priori/posteriori distribution (Guan, 2010). Some of the

other ability parameter estimation methods that exist in the literature are Weighted Likelihood Estimation (WLE) (Warm, 1989), Maximum a Posteriori (MAP) (Samejima, 1969), logistic regression (Reynolds et al., 1994).

Item selection is another important component of CAT. In CAT, items are selected in real time, and the choice of an item cannot be undone once it is administered. For this reason, item selection, which is sequential by nature, is another important component of CAT (van der Linden, 2005). There are various item selection methods used in CAT each of which has its own strengths and limitations. The most prevalent methods are Maximum Fisher Information (Fisher, 1922), a-stratification (Chang & Ying, 1999), Likelihood Weight Information Criterion (Veerkamp & Berger, 1997), Gradual Information Ratio (Han, 2009) and Kullback-Leibler (Chang & Ying, 1996).

Another important decision in CAT design is stopping rules (Wainer et al., 2000). Two of the most commonly used stopping rules are fixed length and variable length. A common varying-length rule in CAT is to stop the test when the standard error (SE) of ability estimate falls below a threshold (e.g., $SE < 0.30$; Morris et al., 2020). This termination rule ensures a desired level of measurement precision for all examinees on the condition that the distribution of item locations in the item bank are uniform (Babcock & Weiss, 2012; Wainer et al., 2000). When an item bank has limited depth at certain ability levels, the SE stopping rule can be inefficient (Choi et al., 2010). If there is an insufficient number of items that give information at a specific ability level, an excessive number of items may be needed to reach the predetermined SE threshold. Another common stopping rule is fixed length stopping rule. Under a fixed length stopping rule, CAT is terminated when a predetermined number of items have been administered. The primary advantage of the fixed length

stopping rule is its simplicity and similarity to PBT (Gushta, 2003). It has even argued that variable-length CATs result in poorer or biased measurement than fixed-length CATs (Stocking, 1987; Yi et al., 2001). However, one limitation of the fixed length stopping rule is that examinees are measured with varying levels of precision and with larger measurement error, particularly at extreme trait levels (Choi et al., 2010).

To summarise this section, while there are several advantages CAT may provide test administrators and takers, there are also some limitations. The following section discusses a variety of issues that should be taken into consideration when switching from PBT to CAT.

Comparability of Scores across Different Test Delivery Modes

Taking a test in front of a computer as an adaptive mode is unarguably different from taking it in a paper-and-pencil format in many ways. Some of the issues extensively researched are if these differences are critical; for example, do they disadvantage one examinee over the other? Most important is ensuring that CAT scores are comparable to a score an individual would earn using a PBT.

Comparability between the CAT and the existing PBT version of the test is a crucial issue and cannot simply be assumed without statistical and empirical evidences. Comparability gains more importance especially if scores on a CAT and PBT are to be used interchangeably in high stakes testing situations, which is generally the case when the CAT version was first introduced and was not immediately aimed to replace the PBT. At this time, there were concerns about how differences in test questions, test scoring, testing conditions and examinee groups could potentially lead to serious problems with comparability.

As Wang and Kolen suggest (2001), comparability between the CAT and PBT version can be examined through live CAT administrations and simulation procedures based on generated data, real data or a combination of both (hybrid). While simulation methods are generally applied at the developmental stage in which the basic design of the CAT version is determined, live CAT administrations are conducted to check and adjust the CAT version previously developed and tested with simulations. Apart from this, live CAT administrations are primarily done to evaluate the overall comparability in terms of reliability and validity. Simulations, on the other hand, are commonly used to evaluate the psychometric properties of the items and content validity of the test. Wang and Kolen (2001) developed a framework of criteria for evaluating comparability including three categories of criteria: validity criterion, psychometric property/reliability criterion, and statistical assumption/test administration condition criterion. For the purpose of evaluating comparability, the present study benefited from certain elements of the psychometric property/reliability criterion in the framework. After evaluating comparability, a classification analysis was performed to see the extent to which CAT test results accurately classify examinees with reference to PBT.

Classification with CAT

Both for reporting and inferential purposes, “a primary outcome of many assessments is to classify examinees into categories” (Lathrop, 2015, p. 1). To illustrate, examinees can be classified as advanced, proficient, basic, and below basic depending on their test scores. The classification of each examinee depends on some estimate of his or her ability which includes some amount of measurement error. In order to do a classification analysis, an examinee’s score is compared against some

cut-off value. The purpose is to classify examinees as accurately as possible with maximum 5% error rate” (Lathrop, 2015).

Traditional PBTs are usually long enough including many items to prevent examinees’ misclassification. Moreover, PBTs are constructed to measure around the mean level of examinees traits and are less sensitive to deviations from the mean. For this reason, CAT has been suggested to classify examinees from various ability strata with much less items and higher degree of precision (Davey & Pitoniak, 2006; Wainer et al., 2000; Way et al., 2006).

To make classification decisions, an examinee’s score is compared against a cutoff point. The purpose is to classify the examinee as accurately as possible with no more than a 5% error rate. In CAT context, both the estimated theta and its associated standard error of measurement (SEM) are employed. An examinee can be classified as being above a cutoff point on the ability scale provided that the estimated theta and its 95% confidence interval, which is calculated as plus or minus two SEMs, are above or below the cutoff point (Weiss, 2004). The test can be terminated when this condition is satisfied as CAT can evaluate this decision after each item is administered. The outcome of such a test will be “a set of classifications made for a group of examinees that all have at most a 5% error rate” which can be controlled “by changing the size of the SEM confidence interval around θ ” (Weiss, 2004, p. 75). CATs may have its own caveats though. The first limitation involves content balancing as some traits might include two or more subdomains. Besides, many items should be used to make repeated measures at that specific level possible without overlap between examinees, which necessitates a large bank of items accessible. Item security is another limitation of CATs since some items may be

chosen more frequently than other items or passed on to the other examinees (Wainer & Eignor, 2000).

In the present study, classification of typical CAT is evaluated based on two important indices: classification accuracy (CA) and classification consistency (CC). The study of CA and CC plays a key role in improving the design of assessments as well as contributing to public understanding and confidence in those assessments. CA and CC are two indices indicating the quality of the classification decision. CA estimates “the extent to which actual classifications using observed cut scores agree with true classifications based on known true cut scores” and is closely related to the validity of the classification (Lee, 2010, p. 1). On the other hand, CC estimates “the rate at which the classification decision will be the same on two identical and independent administrations of the test, and so has a strong relationship to the reliability of the classification” (Lathrop, 2015, p. 1)

In the context of CAT, discussion on classification has been more focused on CATs which are specifically designed to classify individuals (Eggen & Straetmans, 2000; Gnams, & Batinic, 2011; Spray & Reckase, 1994; Rudner & Guo, 2011; Weissman, 2007; Yang et al., 2006). Such tests are called computerized adaptive classification tests (CACT; Thompson, 2007). CACTs generally use varying-length CATs and then a probabilistic approach to make a classification decision. Many adaptive testing programs are fixed-length and provide a score used solely or along with other criteria to make decisions on examinees. (Cheng & Morgan, 2012). Although there are various testing programs where regular fixed-length CAT is used to classify examinees, discussion of classification performance for regular CATs is lacking in the literature.

Problem

Despite the worldwide popularity of CAT, it is important to ensure CAT is comparable to conventional PBTs. This comparison is even more critical when a testing center is transitioning from PBT to CAT and both exams may be being administered to student populations. Comparability between the alternative test versions cannot be assumed and empirical evidence need to be obtained to ensure that the examinees are treated fairly across alternate testing modes (Wang & Shin, 2010, p. 5). As Buerger et al. (2016) indicate the comparability of computer-based and paper-based tests is essential because it is a prerequisite for comparing computerized scores with scores from previous cycles. Evidence of comparability between different administration modes needs to be obtained and provided to ensure conditions that lead to stable large-scale assessments (Mazzeo & von Davier, 2013).

When a PBT test is administered in CAT format, the medium of administration turns from paper to computer and the test delivery algorithm changes from linear to adaptive. In adaptive testing, test delivery paradigm enables the test items to be chosen and administered to tailor them to each examinee's ability. For this reason, administration mode and paradigm effect should be examined both so as to ensure the CAT and its paper-based format can be compared. Although the administration mode has been widely researched in the comparison of paper-based and linear computer-based versions of a test (Bennett, 2003; Choi et al., 2003; Dunkel, 1991), the situation gets more complex when comparing CAT with its paper-based version.

A study which aims to examine CAT version of a language, or any test in general, proficiency test should start with comparability of scores estimated by PBT and CAT. Comparability between scores from different test delivery modes is more

important when a test is concurrently offered in CAT version and its conventional PBT version. Likewise, understanding the classification performance of assessments is crucial where key decisions are taken based on the grades or classifications.

The second problem this dissertation focuses on is the classification made by CAT. One of the common purposes of educational measurements is to inform the classification of individuals for various purposes such as professional licensure, admissions to educational courses of study, and migration. Classification decisions are often high stakes as classification errors may cause examinees to lose opportunities such as well-deserved educational development (Zhang, 2010). Despite this, misclassification of individuals is common and may result from measurement errors associated with sampling, equating, and the assignment of cut scores, standard setting methods and standard setting committees (Arce-Ferrer et al., 2002; Linn, 2003; Stone et al., 2005). Classification of individuals that record scores on assessments close to decision cut-off points is particularly critical because misclassification is highly likely and may result in inappropriate decisions (Eckes, 2017). Given the significant impact classification decisions may have on individuals' lives, it is incumbent upon test developers to provide evidence of classification accuracy and consistency (Lathrop, 2015). Reporting information about test classification performance is also recommended by AERA, APA and NCME, 2014. Not every CAT is designed specifically for classification purposes with varying-length format. Instead, there are many CATs using fixed number of items to classify individuals. CATs such as ACCUPLACER (College Board, 2003) produces estimates using fixed-length approaches. Thus, classification performance of non-CATs needs to be examined prior to making such decisions to ensure validity and reliability of classification.

In brief, the potential for CAT to improve classification decisions for the purpose of placement into language programs demonstrates that this type of assessment is worth exploring in the context of English medium university admissions testing. This is crucial because, in addition to binary pass-fail classification, classifying individuals into more than two groups for placement purposes is a commonly expected function of locally developed language tests (Dimova et al., 2020). Although the literature review indicates that a CAT version of the English language proficiency test may improve classification consistency and accuracy for admissions and placement decisions, questions remain regarding test termination rules and the appropriate number of cut-off points.

Purpose

The present study was conducted to compare several psychometric qualities of the CATs and investigate the classification performance of CATs across various test termination rules and the use of a single cut-off score at different locations of ability continuum versus multiple cut-off scores. The analyses was performed on the reading subtest of a language proficiency test and the results were compared against the PBT version of the subtest.

The study has two stages. In the first stage, comparability of CAT was examined under different testing strategies in the context of English language proficiency assessment. Comparable results produced by CAT with PBT will be considered as psychometric evidence for applicability of CAT for English language assessment. To this end, a series of simulations were conducted, and statistics produced by CAT and PBT were compared through statistical analyses. Number of items in CATs, SE values, ability estimates and correlations were compared against PBT.

In the second stage, classification performance of CAT was investigated across various CAT scenarios defined by different test termination rules, and the use of one cut-off score versus multiple cut-off scores, which communicate a lot of information as to the quality and utility of language tests. Classification performance of CAT was explored by two ways. First, number of examinees who are classified by PBT and CATs in the same or near percentiles were examined. Then classification performance was investigated by CA and CC values for different CAT scenarios in a comparative manner with PBT.

Research Questions

To address the gap in the literature, the research has the following research questions:

1. To what extent do fixed-length and SE-based CAT administrations of English proficiency reading subtest produce ability scores comparable to PBT?
 - 1.1. Is there a difference between CATs and PBT in terms of test difficulty?
 - 1.2. Is there a difference between CATs and PBT in terms of the number of items administered?
 - 1.3. Is there a difference between CATs and PBT ability estimates in terms of standard error?
 - 1.4. What are the correlations between CATs and PBT ability estimates?
2. What is the classification performance of fixed-length and SE-based CAT administrations of English proficiency reading subtest when compared to PBT?

- 2.1. How does CAT classify examinees into different ability strata compared to PBT?
- 2.2. How do CA and CC values differ between PBT and CAT versions of the subtest simulated with different test termination rules?
- 2.3. What is the effect of the location of the cut-off point on CA and CC values when a binary pass-fail decision is required?
- 2.4. How do CA and CC values vary when multiple cut-off points are used?

Significance

Along with the shifting perspectives and reform movements in education, it is now easier to see that the underlying testing paradigm in most assessment programs will inevitably be scrutinized and more flexible assessments systems are likely to be in place. To this end, new technologies will be more commonly incorporated into standardized testing to keep up with the changes and meet the demands. CAT is one of the innovations that will draw more attention in this sense both nationally and internationally.

Research on CAT in language assessment usually involves the development of a computerized version of a test and comparison with a conventional paper-based version of the same test. Mizumoto et al. (2019) report on the development of the CAT version of the Word Part Levels test and conclude that CAT decreases the time required to complete the assessment and leads to higher measurement accuracy. Increased measurement accuracy has also been reported for CATs of vocabulary size (Tseng, 2016), and reading and listening proficiency (He & Min, 2017). The potential for CAT to increase measurement accuracy in situations where the test is

designed to assess reliably at different levels of ability, for example for purposes of placement testing, is clear but relatively unexplored in language testing.

While the original purpose of CATs was to obtain a precise estimate of an examinee's ability, "CATs have also shown to be useful in classification problems" (Eggen & Straetmans, 2000, p. 713). The researchers described CATs for contexts whereby the primary goal is in classifying the examinee into binary categories like pass/fail or master/nonmaster, not in estimating the ability of an examinee (Lewis & Sheenan, 1990; Spray & Reckase, 1994, Weiss & Kingsbury, 1984). According to Diao and Sireci (2018), "in many modern testing programs, it is the classification decision, rather than a total test score, that carries the highest stakes" (p. 24). On a similar note, Bachman and Palmer (1996) state that test scores are primarily used for decision making. These decisions are considered high-stakes as they can have a significant impact on individuals and institutions especially when test scores are used to inform admissions or licensure decisions. Therefore, it is inevitable to ensure that decisions made are fair and valid for this key group of examinees.

As stated earlier, CACTs have an extensive literature. However, more studies are needed on classification performance of regular CATs. This study is expected to fill this gap by examining classification issue of a CAT on a real data set. On a related note, estimating classification performance provides an important source of information for evaluating the technical quality and utility of contemporary educational assessment. Although the need to investigate classification performance is important for supporting the use of a test for a particular purpose, methods for doing so and factors affecting classification such as cut score are not widely discussed in the literature, especially in the field of language assessment.

CHAPTER 2: REVIEW OF RELATED LITERATURE

Introduction

This chapter establishes the theoretical framework within which the current dissertation was conducted by providing a review of classical test theory, item response theory, computer adaptive testing, language testing and classification including the research done on each field.

Classical Test Theory (CTT)

CTT is the approach that most individuals have been exposed throughout their education. CTT refers to a psychometric framework for constructing tests and analyzing test performance (Sijtsma, 2015). CTT relies largely on correlational concepts. CTT assumes that an observed score equals the true score plus error ($X = T + E$) and reliability equals 1 minus error ($r = 1 - E$). In this theory, the correlation coefficient between true scores and observed scores is called the reliability index and the square value of this index indicates the reliability coefficient and generally Pearson r is used for this purpose (Helms, 1999).

The concepts of error and true scores are crucial in CTT. CTT involves a set of assumptions regarding the connection between true or observed test scores and the factors that influence these scores, which are usually referred to as error (ALTE Authoring Group, 1998). The basic idea of CTT is that measurement suffers from random measurement error, referring to the difference between a test score obtained on one testing occasion and the true score. Random measurement error may result from unpredictable changes in a person's mood, concentration, susceptibility to

external stimuli, or physical condition and it negatively affects a person's test performance at an unknown magnitude.

Within the framework of CTT, observed scores, the examinees' observed scores on a test, contain errors (unsystematic effects owing to factors not measured), which consequently lead to error variance (unsystematic variation in the scores resulting from random errors (Fulcher & Davidson, 2012). Error variance is considered random as it can stem from numerous factors. The first factor is extraneous sources which are non-systematic sources that are not relevant to the real purpose of the test. According to Brown (2005), some of the examples of such extraneous sources would be environmental factors like noise, administration procedures like equipment, scoring procedures like scorer biases, test items like item quality and the examinees themselves like fatigue.

The other notions central to CTT involve the notions of strictly parallel items and item variance, as well as parallel tests. The term strictly parallel items refers to the conditions necessary to derive CTT procedures for estimating the reliability and standard error of measurement for a given test (Fulcher & Davidson, 2012). Strictly parallel items must be unidimensional, independent of all other items in terms of error, independent of true scores in terms of error, equally good indicators of true scores and have equal variances. In practice, while piloting or revising language tests, test developers are trying to make sure that the items are designed to test a single construct equally well. Moreover, in order to minimize the degree of correlation between item's errors and with true scores, test development and administration procedures are utilized. This process is followed by maximizing the extent to which all items discriminate similarly, the extent to which they lead to

equal variances and the level to which items are equally good indicators of true scores.

Despite straightforward and practical nature of CTT in testing practice, the theory has some major shortcomings that influence measurement quality. One of the limitations of CTT is that the items that maximize internal consistency reliability are often the items suitable for the average examinees in the group as they match with the difficulty levels of these examinees (Weiss, 2004). When the same conventional measurement instrument is used for individuals whose trait levels deviate from that trait range, it provides poor measurement as the items have little relevance for the trait level of those examinees. The implication of this is that for the examinees deviating in ability from the level of the conventional test, the test will provide very little information (Weiss, 2004). Moreover, time constraints that are often imposed on conventional tests introduce such traits as persistence or slowness which interfere with efficient measurement of the trait that the test is designed to measure. Weiss (2004) also notes that despite the high reliability and low standard error of measurement for that set of items, the decrease in measurement imprecision is assumed to be constant across the measurement scale.

One of the major shortcomings of CTT is that an examinee's ability is defined depending on a certain test (Hambleton et al., 1991). Within the framework of this theory, it is not possible to separate examinee characteristics and test characteristics. Test and item parameters vary as the test context changes and as a natural consequence of this, item parameters and test score are only expressed depending on a particular group of examinees, which means item characteristics are group-dependent and scores representing examinee ability/proficiency are test-dependent. Due to the absence of test-free scores, it is essential to administer the

same items to ensure proper comparisons. It is well known; however, that measurement precision is endorsed when test items are selected in line with the specific ability levels of examinees (Choi, 1992). Apart from these, an indicator of sample or group dependency of item and test statistics in CTT is that the difficulty of a test item is determined by the rate of examinees in a group who answer the items. For example, in classical psychometrics, knowing that the difficulty level, p^1 , of a chemistry item is .60 does not show us the level of difficulty unless we know about the relevant group of examinees. In short, within the framework of CTT, examinee ability levels are test dependent. Furthermore, item and test parameters like reliability coefficients are also sample dependent. The absence of test-free and sample-free item and test statistics in CTT has hindered behavioral measurement from reaching its potential by causing significant practical problems and interpretation difficulties that the model fails to solve.

It has been discussed in the literature that CTT has become outdated or overshadowed by the development of additional theories and techniques, particularly criterion-referenced testing, generalizability theory (G theory) and IRT. However, considering the rare use of criterion-referenced testing, G theory and IRT in language programs all around the world, it would be fair to say that classical does not mean old-fashioned and CTT is still alive. As a matter of fact, CTT is probably the most dominant psychometric theory actually applied to the problems of language testing in real-world language teaching contexts except for few test development programs (Fulcher, 2012). Tejada and Rojas (2005) dwelled on how recent research has been utilized in favor of the use of IRT-based models, rather than CTT. Likewise, it was

¹ In CTT context, p denotes the item's difficulty level. Items with high p values are easy items and those with low p values are difficult items (Kline, 2005). Another widely used parameter in CTT is item discrimination which means the potential of the item to distinguish examinees with a low or high performance on the test according to their total test score and item score (McDonald, 1999).

projected by Harvey and Hammer (1999) that IRT-based methods will mostly replace CTT-based methods over the upcoming years. In his criticism of CTT, Borsboom (2005) suggested that very few researchers in psychology think CTT is an appropriate model to measure psychological constructs. On the other hand, although CTT has been dismissed by some scholars as obsolete and irrelevant, it is also argued that CTT still has uses in modern psychometric issues and will not perish. It has been noted that analysis on classical test theory-based psychometric procedures remain valuable and informative (Casillas et al., 2009). The fact that a certain tool is more powerful than the other one does not necessarily mean that it is always preferable to less powerful tools. In the following section, one of the measurement theories, item response theory, widely considered as a strong alternative for CTT, is presented.

Item Response Theory (IRT)

IRT involves a broad set of statistical models which intend to show the probability that an examinee will choose a particular item response, which depends on item parameters (Immekus et al., 2019). Item parameters involve the item difficulty (b), item discrimination (a) and pseudo-guessing (c). Item discrimination refers to the degree to which an item discriminates among examinees along the underlying ability continuum. Item difficulty refers to the point on the underlying ability continuum where an examinee has a probability of .50 to select a particular response. For instance, the threshold for a dichotomously scored item (i.e., the response is either correct or incorrect) is a measure of item difficulty indicating how easy, or difficult, the examinees found the item. On the other hand, measurement instruments can also consist of ordered-categorical items such as Likert scale. The threshold for such instruments is the point on the trait scale in which an examinee would have a probability of .50 to select a particular response category. The point

where the examinee stands on the measured trait (e.g., motivation) denotes their ability or theta symbolized as θ^2 (Crocker & Algina, 1986). For multiple-choice items, IRT adds a third item parameter called as pseudo-guessing. It basically refers to the probability that an examinee with a very low ability level will correctly answer an item merely by guessing. Depending on the parameters used, some of the most popular logistic IRT models are called 1PL, 2PL, and 3PL IRT models.

IRT offers important advantages for examining the psychometric features of instruments (Hambleton & Swaminathan, 1985; Reise & Revicki, 2013). The first attractive feature of IRT is that the item parameter estimates are sample-independent or invariant of the sample unlike the item parameters obtained within the framework of CTT (Crocker & Algina, 1986). Therefore, a well-fitted IRT model would yield the same item parameter values regardless of the sample (Embretson & Reise, 2000). Item parameter invariance is an important principle of IRT underlying the estimation of an individual's probability of a correct response at any given ability level (Immekus et al., 2019). On a related note, an individual's estimate on the measured trait is independent of the items administered, which is called invariance of ability estimates. As a result, after a set of items has been calibrated, an individual's ability estimate can be obtained using any selected subset of items. Therefore, it is not necessary to administer the same set of items to estimate an examinee's ability estimate. Instead, a sample of items can be selected from a bank of calibrated items and administered to examinees to identify their location on the measured trait, which constitutes the basis of CAT. Lastly, unlike CTT which considers the standard error of measurement as the same for the entire score distribution, IRT provides a measure of the precision, or standard error, for each trait estimate (Crocker & Algina, 1986;

²Theta (θ), in IRT, refers to an unobserved hypothetical variable (a latent trait) such as arithmetic or linguistic ability (Baker & Kim, 2004, p. 3.)

Harvill, 1991). As with other statistical procedures, such assumptions has to be checked and empirical evidence must be collected to guide model selection decisions and researchers need to familiarize themselves with the available data and models.

It is commonly argued that conventional PBT is not efficient because the most items do not give useful information in distinguishing among examinees deviating from particular ranges of the latent trait. On a typical conventional exam, examinees with a high trait level see easy items which provide very little information about relative trait standing. Similarly, examinees with a low trait level may receive items which are too difficult and hence, provide almost no information about their trait level. IRT-based CAT attempts to overcome this inefficiency “by attempting to administer each examinee a test for which their proportion correct will be around .50” (Embretson & Reise, 2000, p.263). The purpose of testing has been commonly believed to be precise measurement across the entire ability range. This entails an item bank with sufficient number of highly discriminating items and with difficulty parameters covering the ability range. (Weiss, 1982). According to Embretson and Reise (2000), an item bank meeting these criteria results in an appropriately tailored test and enables all examinees to be measured precisely regardless of their location on the latent trait. The primary contribution of IRT is that it allows for locating items and examinees on the same scale and tailoring the difficulty of the test to the ability level of the examinee although different sets of questions are administered to the examinees (Sands et al., 1997).

Unlike CTT, IRT operates with probabilistic models for positioning each examinee’s ability and item parameters (difficulty, discrimination and guessing) along a hypothetical ability continuum and testing goodness-of-fit to model predictions (Choi, 1992). In IRT, it is assumed that an underlying ability continuum

with the least amount of ability is located in the far left and the highest ability in the far right. The ability of every examinee has a place across this continuum. As indicated by Fulcher and Davidson (2012), average ability is set to 0 and located in the middle of the continuum. In general, most examinees' ability estimates range between -3 (low ability) and +3 (high ability) on this scale for a typical normally distributed trait. The main principle underlying IRT models is to carry out test analysis not based on arbitrarily defined test scores but on the item level.

The commonly IRT models are 1PL (also known as Rasch model), 2PL and 3PL models. Rasch was the first to develop one-parameter logistic model, also known as Rasch model with the assumption that the success of an examinee in a particular item should depend on the product of his or her ability, denoted by theta (θ) and item difficulty (van der Linden & Hambleton, 1996). One of the fundamental postulates of IRT is that both persons and items are located on the same continuum, which enables to make comparative statements about how a typical person might respond to an item (De Ayala, 2009)

The probability of a correct response to an item is determined by the ability level of the examinee and the item parameters (Lu et al., 2012). Unlike Rasch, Lord and Birnbaum held the belief that at least two parameters are required for an IRT model: difficulty and discriminating power of the items. Lord developed the two-parameter normal-ogive model³ for analyzing multiple-choice test data and Birnbaum introduced two- and three-parameter logistic models. In addition to item discrimination (a) and item difficulty (b), pseudo-guessing parameter (c) is also

³Normal ogive model was the first IRT model for measuring a latent trait (Ferguson, 1942; Lawley, 1943; Mosier, 1940, 1941; Richardson, 1936). It was later improved by Lord et al. (1968). In the model, an item characteristic curve (ICC) is obtained from the cumulative density function (CDF) of a normal distribution. In 1950s, Rasch developed a less sophisticated IRT approach in the 1950s using a logistic function to obtain an ICC instead of the normal ogive function (Rasch, 1960), simplifying the normal ogive model and the computation required.

incorporated into 3PL model. Pseudo-guessing or pseudo-chance parameter (c) is a parameter for guessing and it is used mostly for multiple-choice items. 2PL model does not include the pseudo-guessing parameter (i.e. $c_j \equiv 0$) and 1PL model uses only the b_j coefficient (e.g. setting $a_j \equiv 1$). One of these models need to be chosen depending on the data. While models are mostly used in psychology to connect behavior to constructs or experimental manipulations, the term *model* in psychological measurement refers to a mathematical model in which independent variables are combined numerically to estimate a dependent variable (Embretson & Reise, 2000). Drawing valid inferences from IRT models requires a good fit of the data to the model and violations of model-data fit limit the usefulness and applicability of the model (Ames & Penfield, 2015).

There are several methods used to assess the goodness-of-fit of particular IRT models. Two general approaches to evaluating how well the IRT model explains or predicts the responses to a particular item (i.e., item fit) are available in IRT framework. Embretson and Reise (2000) describe graphical procedures and formal statistical procedures that assess item-fit. The former judges item-fit on the basis of a comparison between an estimated item response curve (IRC) and ‘empirical’ IRC without performing any statistical tests (p. 234). Some of the causes behind problems in item fit may be due to “a failure to estimate enough item parameters (e.g., when a 1PL is fit to 2PL data)” or “poor item construction”. (Embretson & Reise, 2000, p.234). There is a substantial amount of work on IRT model data fit analysis. Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999) require evidence of model fit if inferences are to be made from a data set with an IRT model. To evaluate model fit, item fit plots are drawn, and item fit statistics are calculated through commercial software packages

such as PARSCALE (Muraki & Bock, 1991). Sinharay et al. (2011) emphasized that Type I error rate for the item fit statistics available in commercial software packages leads to some items being wrongly labeled as misfitting and removed from the item pool. Besides, they argue that the judgment on model fit relies heavily on the person examining the plots and therefore is subjective.

On a final note, it would be nice to reiterate that “no model is ever a perfect reflection of the data it is to summarize” (Wainer & Thissen, 1987, p.339). In other words, there will always be errors of fit and the same possibility exists with IRT models. It is; however, important to know about the degree to which measurement accuracy in a particular model is affected by misfit and what can be done to minimize it. It is stated in the literature that IRT is based on stronger, or more restrictive assumptions than is CTT, and therefore it is able to make more accurate inferences about examinees’ performance on each item, their levels of ability, and about the characteristics of test items (Bachman, 1990).

Item and Test Information

Unlike a conventional test, the concepts of test and item information are central to IRT methods rather than the notion of internal consistency. The term information is a statistical indicator of the quality of a parameter estimate (Reckase, 2009) and it is used to describe to what extent we feel certain about the estimate of a person’s location on the ability continuum. The concept of standard error, on the other hand can be used to describe our uncertainty about an estimate.

A prominent feature of IRT is the ability to investigate an item's functioning graphically. Item information curves (IIC) are plotted to this end. ICCs model the probability of an item response for a given ability, or trait, level (Embretson & Reise, 2000; Yen & Fitzpatrick, 2008). In other words, IICs denote the relationship between

the probability of an examinee correctly answering each item and the examinee's ability (Fulcher & Davidson, 2012). The concept of item information is critical to find out how much we can learn about the latent trait based on each item and how each item contributes to the test information (Le, 2013). In item response theory, the item information function is a measure of how much statistical information a test item provides. Item information is a function of θ , the ability, proficiency, skill, or trait measured based on the examinee responses to the test items. The amount of information provided by an item, relative to θ , is closely related to the parameters of the item.

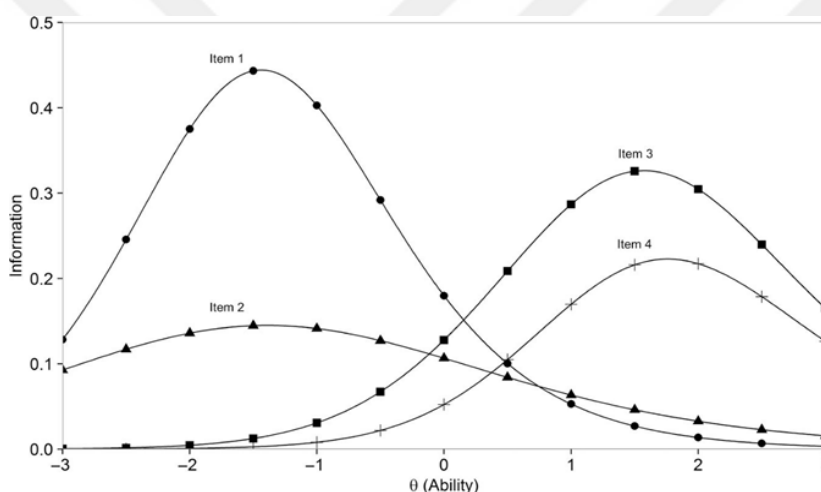
Figure 2 (DeMars, 2018) denotes the information functions for four dichotomous items, scored 0 or 1. It is seen in the figure that items 1 and 2 are more informative at lower theta values, and items 3 and 4 are most informative at higher ability values. Items 1 and 2 have the same difficulty but item 1 is more discriminating. Besides, items 3 and 4 are more difficult than items 1 and 2. If an item has lots of information near an examinee's θ value, the examinee's response to that item helps estimate θ better. In other words, if an item has a high discrimination parameter and difficulty parameter near the examinee's ability estimation, it is useful for estimating θ . If the item is too hard or too easy, or simply not very discriminating, the response to that item does not provide enough information about what the examinee knows or can do.

In Figure 2 (DeMars, 2018), maximum information is reached at $\theta = -1.44$ for item 1, -1.39 for Item 2, 1.57 for item 3, and 1.76 for item 4. The information function steepens as the discrimination parameter (a) increases. Comparing two items with the same item difficulty (b) and pseudo-guessing (c) parameters, the item with the higher a parameter will have more information but in a narrower range of θ .

In Figure 2, b and c values are -1.5 and .05 for items 1 and 2. However, discrimination value is 1.4 for item 1 compared to 0.8 for item 2, which means item 1 has more information for low θ values, but slightly less information than item 2 for high θ values. The c parameter diminishes the information function, especially for hard items. Items 3 and 4 have the same a and b parameters ($a = 1.2$, $b = 1.5$), but the c parameter is .05 for Item 3 but .25 for item 4, which means item 4 provides less information.

Figure 2

Four Item Characteristic Curves



The knowledge gained from IICs about how each item will behave in estimating an examinee's location allows the instrument to be refined or rescaled (De Ayala, 2009). In CAT sessions, item information function is used for (a) calculated standard errors (SE)⁴ for finding test reliability and (b) selecting next item to be given to examinees. Each item in an information curve denotes how well each item performs as a function of the trait. Thus, a high information value is closely linked with a small standard error of measurement. (Le, 2013). An ICC can be obtained for

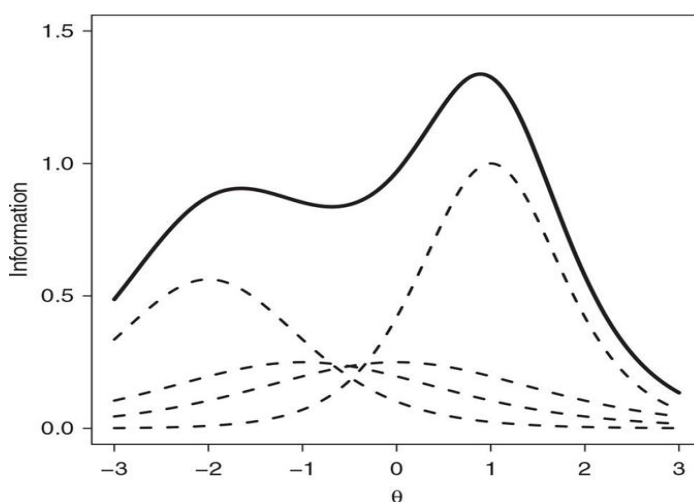
⁴Standard error is defined by De Ayala (2009) as an index of the variability (i.e., the standard deviation) of an estimator with respect to the parameter it is estimating” and the larger the value of a standard error, the less certain we are about the parameter's value or vice versa.

every item to show how measurement precision can change across various trait levels.

Test information function (TIF) is another crucial concept for test development in IRT framework. TIF is a measure of the amount of information provided by responses given on a test about the measured trait. Under IRT, assumption of local independence is the sum of the item information functions (IIFs). Figure 3 (Johnson, 2018) shows IIFs (dashed curves) and TIF (solid curve) for five items with discrimination parameters equal to 1.5, 1.0, 1.0, 2.0 and difficulty parameters equal to -2.0, -1.0, 0.0, 1.0. As illustrated in Figure 3, TIF shows the locations where test is powerful and the locations where test is not powerful. In other words, TIF denotes the locations where the test measures precisely and the locations where the test fails to measure very precisely.

Figure 3

Test Information Function



Measurement precision is not constant with respect to θ since the information changes as a function of θ , which contrasts with CTT as it produces precision measures that are constant with respect to the unknown ability measure θ (Johnson, 2018).

Ability Estimation in IRT

There are various approaches used to estimate θ within IRT framework. Among the most commonly used ones are maximum likelihood estimation (MLE) and Bayesian sequential estimation methods such as expected a posteriori (EAP) and maximum a posteriori (MAP) (Chen et al., 1998; Gorin et al., 2005; Wang & Vispoel, 2005).

One of the common methods of ability estimation is called maximum likelihood estimation (MLE). MLE can identify far more differences among examinees compared to number-correct scoring based on examinee response patterns and item characteristics. For instance, number-correct scoring of a 20-item conventional test can yield maximum 11 different scores (0 to 10). On the other hand, MLE can result in 1,024 (2 raised to the 10th power) ability estimates for the same test (Embretson & Reise, 2000). Unlike, number-correct scoring, MLE also gives an individualized standard error of measurement (SEM) for each examinee. The SEMs from IRT can vary for each examinee depending on how they responded to a set of items. Finally, the theta estimates and their SEMs in IRT can be obtained from any subset of items an examinee has taken as long as item parameters have been estimated on the same scale. MLE utilizes information on the examinee's response patterns and item characteristics and can provide many more distinctions among examinees than can number-correct scoring (Embretson & Reise, 2000). On a final note, MLE does not work if the examinees does not have one correct and one incorrect answer.

EAP, as a Bayesian method, does not require the assumption of normality and mixed iterative mathematical calculations at every stage of the estimations. Unlike MLE, EAP also performs ability estimation when the examinees give all correct or

all incorrect responses (Embretson & Reise, 2000; Hambleton et al., 1991). Because such patterns are likely to occur for the first few items, it is not possible to use MLE for the ability estimation at the beginning of the test (van der Linden & Pashley, 2000). Under the Bayesian paradigm, the posterior distribution of the ability levels (i.e., θ) is defined as the product of the likelihood function and the prior ability distribution. EAP utilizes the mean of the posterior distribution as the θ estimate (Ho, 2010). Choosing a prior distribution for the ability or proficiency level is crucial to Bayesian estimators. The most common prior distribution is the standard normal distribution in which the estimated proficiency levels get smaller toward the prior mean value, leading to relatively smaller overall errors. van der Linden and Pashley (2000) also propose some measures to solve this problem. The first measure suggested is postponing ability estimation until a larger set of items has been answered. Another measure given is using Bayesian methods such as the EAP estimator. A final measure is obtaining initial ability estimates based on empirical information available like scores on earlier related tests.

The primary advantage to the Bayesian methods is that they work when an examinee has only correct or only incorrect responses. Lord (1986) in his discussion of MLE and Bayesian estimation methods from the perspective of IRT, highlights that Bayesian parameter estimation techniques work better than MLE as Bayesian techniques provide more information. Besides, since Bayesian parameter estimation uses psychometric properties of the sample, the same response pattern given by examinees may produce different ability estimations unlike MLE which generates always the same ability estimation for the same response pattern. Lord also proposed the use of Bayesian techniques as they do not lead to divergence. Bayesian methods involve using prior information about the ability distribution of the examinees in the

estimation procedure. Bayes' theorem lies at the heart of the Bayesian procedures. The essence of Bayes' theorem is that both the current data and prior knowledge determines the probability of a given hypothesis. In other words, new information is obtained for a subsequent event and this information is used to revise the probability of the initial event (Triola, 2007). In this context, the terms *prior probability* and *posterior probability* are generally used. As Triola (2007) states, a prior probability is an initial probability value originally obtained before any additional information is obtained. Kline (2005) states that, in Bayesian estimations, the likelihood function is calculated by multiplying a prior distribution to obtain the new posteriori distribution, which is a probability value that has been revised based on the additional information that is obtained later.

Assumptions of IRT

Unidimensionality, local independence and item/ability invariance are three fundamental assumptions that underpin IRT. Viability of these assumptions cannot be assessed directly but it is possible to do it by gathering some indirect evidence to judge the overall fit of the model to the relevant test data.

The first assumption which is regarded as an assumption that cannot be fully met is the assumption of unidimensionality. Most IRT models assume that only one latent trait is needed to account for item responses (Ostini et al., 2015). This means that the items measure a single underlying latent trait and it affects the item responses for each examinee. Other factors influencing the responses are considered as random error (DeMars, 2010). If the assumption does not hold, parameter estimates and standard errors may be questioned. There are more complex models in which it is assumed that more than one ability is essential to explain examinee test performance and these are multidimensional item response theory (MIRT) models,

which are *compensatory* and *noncompensatory* (Reckase, 2009). MIRT embodies a wide range of probabilistic models designed to estimate an examinee's probability of an item response based on item parameters and multiple latent traits (Immekus et al., 2019). Unidimensional IRT models and MIRT models make different assumptions about the latent dimensions. However, it is not easy to choose particular model as there are not enough reliable statistical tests for dimensionality (De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Hattie, 1985; Hattie et al., 1996).

Unidimensionality is a controversial concept because various cognitive, personality, and test-taking factors like test anxiety operate intricately and affect test performance, it is not possible to meet this assumption completely. Language testing, for instance, works on one of the complex components of cognitive processes and it is impossible to meet this assumption strictly in such a test. In language testing, factors, other than the trait to be measured like the testing environment, the test rubric, the language input, the test method and secondary constructs required for each item affect the items responses of every examinee, be it systematically or randomly. The assumption of unidimensionality is presumed untenable as it is argued that there is no item which purely measures any unitary ability (Choi & Bachman, 1992). IRT models face the possibility of a philosophical dilemma regarding the assumption of unidimensionality. According to Hambleton et al. (1991), unidimensionality requires the presence of a dominant ability⁵ that influences test performance. In other words, rather than the assumption of a strict unidimensionality, essential unidimensionality or the presence of a single dominant ability is required as it is less constraining and more likely to satisfy. If an IRT model assumes that a single dominant ability is adequate to account for examinee performance, then the

⁵It must be noted that ability is not always inherent or permanent and ability scores may change over time due to such factors as learning or forgetting.

model is called unidimensional. If; however, more than one ability is required to explain the examinee performance, the model is multidimensional. Multidimensional models have a more sophisticated structure but are less commonly used than unidimensional ones.

Local (or conditional) independence embraces but goes beyond the notion of unidimensionality. It indicates that the response of an examinee to one item will be independent of his or her response to other items, conditional on the latent trait (Le, 2013). In statistics, independence in data occurs, for example, if the outcome of a head in a coin toss has no influence over the outcome of the next toss (e.g. it does not change the probability that the next toss produces a tail). As Linacre (2009) remarks “local independence specifies that the value of one datum has no influence on another once the underlying variable, the latent trait or dimension has been accounted for (conditioned out)” (p. 2). In other words, after considering examinees’ abilities, the assumption of local independence means that an examinee’s response to an item in a test is statistically independent of other responses (Choi, 1992). The departure from this assumption affects the accuracy of item parameter estimates and the total instrument information. Possible reasons behind the violation of conditional independence assumption are listed by De Ayala (2009) as physical dependence among items, content clues, instrument length, insufficient allotted time to complete an instrument (i.e. speediness), and/or an insufficient number of latent variables in the IRT model. Overall, according to this assumption, there is no relationship between examinees’ responses to different items and the set of abilities specified in the model are the only factors which have an influence over examinees’ responses to test items. In IRT models, the probability that an examinee will give a specific response to an item is a function of two components, the examinee’s ability and item

parameters (i.e., difficulty, discrimination, and pseudo-chance parameters). The complete latent space, which represents the set of abilities in the model, consists of only one ability if the assumption of unidimensionality holds. (Hambleton et al., 1991). A crucial point here is that the assumption of unidimensionality and local independence can be regarded as two equivalent properties because local independence is a prerequisite for the assumption of unidimensionality. However, local independence can hold even when the data is not unidimensional. As long as all the traits having an influence over performance are specified, local independence will be obtained. Various testing factors may influence the occurrence of violations of local independence. To exemplify, in a test where a bundle of questions is linked to a single passage, responses on latter items may be influenced by responses on earlier items (Embretson & Reise, 2000). Several statistics have been suggested to give information about local dependence. These statistics are the Pearson's χ^2 , the G^2 statistic, Q_3 coefficient, the Standardized Phi Coefficient Difference, and the Standardized Log-Odds Ratio Difference. When a model shows misfit using an overall goodness-of-fit test, and local dependence is the suspected cause, the general tendency is to locate the source of misfit to take remedial action such as removing certain items or modifying the model. Finally, it is important to note that all the model assumptions are tested simultaneously when using an overall goodness-of-fit statistic (Liu & Maydeu-Olivares, 2013).

Another distinctive asset of IRT is the invariance assumption of item and ability parameters. Invariance often means that values of IRT item parameters need to be identical across different samples of examinees and measurement conditions (Hambleton et al., 1991; Linn, 1990; Lord, 1980; Rupp & Zumbo, 2006. According to this assumption, providing that the IRT model fits the data, parameters will not

vary across administrations, except for sampling fluctuations that cause random error in examinee responses. The property of invariance is regarded as the cornerstone of IRT as it facilitates crucial applications such as equating, item banking, investigation of item bias, and adaptive testing (Hambleton et al., 1991). A primary criticism of CTT is that item statistics obtained with CTT are dependent on the sample of examinees. Analysis based on CTT has been used for a long time and is still useful in test construction, although the trend is definitely towards IRT that provides for measurement independent of a particular set of examinees and items. The assumption of invariance of ability and item parameters is what distinguishes IRT from CTT (Hambleton, 1994). This assumption ensures that item and test statistics remain independent of examinee characteristics. With invariance, it is possible to create optimal individual scores, to tailor tests and to examine test validity by utilizing a wider range of item and score statistics. IRT scoring minimizes score bias and allows for the comparison of scores across different tests.

Limitations of IRT

Despite some benefits of IRT over CTT and the popularity of IRT, its methodology brings a number of unique challenges. First of all, the models are technically more sophisticated, and the parameter estimation methods are generally complicated numerical methods. In addition to item parameters, latent traits can also be difficult to interpret both graphically and numerically. In brief, limitations of IRT can be outlined as follows (Le, 2013):

- Model specifications in IRT are more complicated than CTT.
- A number of IRT graphical and numerical outputs need to be properly interpreted, especially for more sophisticated models.

- Data processing and preparation for use with IRT software can be challenging and time consuming.

In summary, IRT is a family of models that have proved useful in the design, construction, and evaluation of educational and psychological tests (Hambleton et al., 1991). IRT models are especially appropriate for adaptive testing as they make it possible to obtain ability estimates which are independent of a specific set of test items that are administered. It would not be feasible to apply adaptive testing without IRT as it constitutes a framework for comparing the ability estimates of various examinees despite the fact that every examinee is given a different set of items which differ in difficulty. Le (2013) highlighted in her paper that the extensive study of IRT over the years was reflected in the growing number of software packages designed for analyzing item response data from tests. A number of IRT commercial packages have also been created such as BILOG, MULTILOG, WINSTEPS, IRTPRO, MPLUS, and HLM, to name just a few. Furthermore, it is stated that various IRT packages developed in the open source R software to estimate the IRT models have also appeared and gained recognition.

Computerized Adaptive Testing

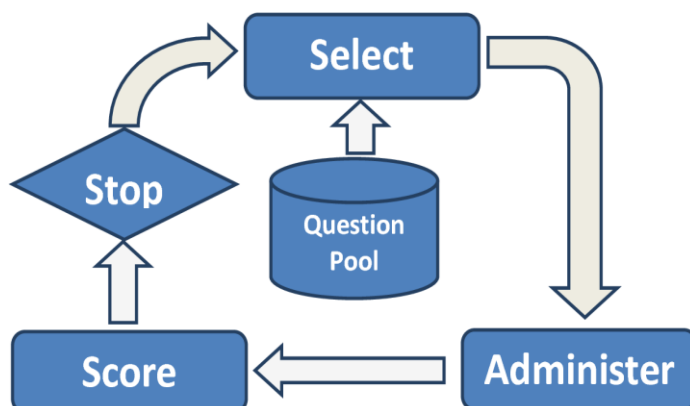
In the history of testing, a tradeoff between individual testing not made up of too many items inappropriate to the examinee's ability level and group testing offering the benefit of uniformity and relatively low-cost has always existed (Wainer et al., 2000). The decision has usually been in favor of group-administered tests. However, a major limitation of mass-administered tests originates from the fact that there is a wide range of ability to be tested. PBT-based group testing forces many examinees to spend time on questions that may be either too easy or difficult for them by requiring them to take the same set of items. This contrasts with CAT which

allows examinees to answer only questions suited to their abilities by selecting the next item to be presented to an examinee after each item response based on his performance (Sands et al., 1997).

CAT was originally developed as an individualized testing system focusing on the efficient estimation of the ability of a student (Lord, 1970; van der Linden & Glas, 2010; Wainer et al., 2000). The basic assumption in CAT is that easy items for high-ability examinees and difficult items for low-ability examinees give little or no information about the examinee's ability. Using an IRT-calibrated item bank, a CAT algorithm ensures that each examinee receives an optimal test. The algorithm selects items from the item bank tailored to the ability of the examinee, based on the previous responses given during the test. Since the questions are tailored to an individual's ability level, CAT naturally offers many advantages for examinees and evaluators which are mainly associated with efficiency and a better user test experience. The order of items in a CAT depends on the examinee's performance on previous items (Hambleton et al., 1991). The common practice in CAT is that an examinee answers two or three items and an initial ability estimate is derived based on his responses to these questions. Later on, items which give maximum information (items are selected using an item selecting procedure) about the examinee's ability are administered in order to shorten the test without sacrificing measurement precision. A computer is programmed to choose from a set of items which will give the maximum information about the examinee's ability. The process of administering items to the examinee continues till a pre-defined test termination rule. A schematic representation of the testing algorithm governing CATs is given below in Figure 4 (Davey, 2011).

Figure 4

Schematic Representation of a Computerized Adaptive Test



Note. From “A guide to computer adaptive testing systems educational testing service for technical issues in large-scale assessment (TILSA)” by T. Davey, 2011, State Collaborative on Assessment and Student Standards (SCASS), p. 5.

<https://files.eric.ed.gov/fulltext/ED543317.pdf>

CATs presuppose the availability of a calibrated item bank with a collection of items constructed to measure a well-defined construct (Eggen, 2018). The item bank also contains various characteristics of each item involving the item parameters derived from the calibration, or the estimation and establishment of model fit with an IRT model. In order for items to be calibrated, they go through certain procedures such as pre-test items. However, the optimal minimum calibration sample size for pre-test item estimation in CAT is one of the issues that need to be considered by CAT practitioners. One of the recent studies that aimed to investigate this issue was conducted by Şahin and Weiss (2015). The findings of the study illustrate that θ estimates in CAT are robust against fluctuations in item parameter estimation and item pool size if the item bank consists of items with high information in line with the examinees' ability levels. Unlike the findings of the prior research, it was suggested by Şahin and Weiss (2015) that a sample consisting of 150 examinees can

be feasible for item calibration in a CAT, and an item bank of 200 highly informative items would also be useful for different purposes.

Another recent study was conducted by Kalender and Berberoğlu (2017) suggesting a CAT version of the Higher Education Transition Examination (HETE) in Turkey as a viable alternative to overcome the psychometric flaws of the exam. Students' raw scores on the science subtest of the HETE were used in the study. An important aspect of the study is the use of an external criterion. Students' raw scores on the mathematics subtest of the HETE were used for this purpose. Considering my own study, I benefited a lot from reading the study especially in terms of research design and methodology. There are two phases of the study. The first phase involves CAT simulations and the second phase involves administering a live CAT. The results seem to advocate the CAT version of the subtests as a feasible approach to paper and pencil version of HETE.

In conventional tests, examinees are generally given the same items, which means proficient examinees are frequently administered items that are too easy for them, less proficient examinees receive lots of difficult items. These items are not very informative and do not contribute to an individual's proficiency estimate. However, in CAT, an examinee's interim proficiency estimate depends on the choice of the next item. In CATs, each examinee is administered different items depending on their proficiency estimates and they are administered only as many items as required to reach a decision. In this way, CATs generally bring about a significantly reduced test length. CATs has the benefit of increasing the measurement precision as after each item selection, the examinee's ability is updated and the next item selected improves the accuracy for this estimated ability (Yao, 2019). Compared to a conventional test format, a CAT environment has many more delivery and

measurement complexities. One of the complexities is item selection and ability estimation. The theory and practice of item selection and ability estimation for CAT is still evolving. Item selection and ability estimation is considerably harder as the test adapts to the examinee in a CAT. Unlike conventional tests, item selection and ability estimation proceed hand in hand in a CAT (van der Linden & Glas, 2010). Efficient ability estimation largely depends on appropriate selection of items for an individual. Likewise, appropriate selection of items for an examinee heavily relies on the efficiency of ability estimation. Robust procedures are needed to solve the complexities of a CAT format with little or no human intervention. Item selection, ability estimation and test termination rule are the main components of CATs.

Ability Estimation

The criterion used to estimate the ability level of an examinee is an important feature of any adaptive testing strategy. In a CAT, an examinee's ability is estimated after each item or a set of items is administered and scored (Weiss, 2011). IRT ability estimation methods facilitate after a single item or a couple of items is administered. Two major advantages of IRT ability estimation are that it uses the information in an examinee's responses along with the item information available and that every time ability is estimated, it provides a standard error of the ability estimate, which denote the confidence that the test user can have in a given ability estimate and can be used to terminate a CAT for an examinee (Weiss, 2011).

According to van der Linden and Pashley (2000), in CAT, there are three stages of ability estimation: ability estimation to start the item selection, ability estimation during the test to tailor item selection to the examinee's ability and ability estimation at the end of the test to report a score to the examinee. The primary function of final ability estimates is to give the examinee a meaningful summary of

his performance as well as guiding item selection. To this end, as van der Linden and Pashley (2000) indicate, final estimates are sometimes converted into an equated number-correct score on a reference test. To conclude, ideal ability estimation method is closely related to other aspects of the CAT such as the choice of item selection criterion and the composition of the item pool.

Item Selection

The criterion used to select the test items administered to an examinee is another important psychometric feature of a CAT. In a CAT, items are selected, administered, and scored one at a time, θ is estimated after every single item or a set of items is given, and a new item or another set of items is selected to continue the test (Weiss, 2011). When only a single item or two has been administered at the early stages of a CAT, the next item is typically selected by a step rule. If the first item was answered correctly, the examinee's original prior ability estimate is decreased or increased by some amount (e.g., .40) depending on the response given. As the test continues and a response pattern of at least one correct and one incorrect response is obtained for an examinee, MLE is used to obtain a new theta estimate based on the examinee's responses to the items administered up to that point in the test. Item information function plays a key role in item selection in CAT. The item information function denotes the contribution an item makes to the accuracy of the measurement of the student's ability (Eggen, 2012). The item with maximum information at a given ability estimate is selected in CAT, searching for every examinee the items on which he or she has a success probability of 0.50.

A CAT is referred to as an unconstrained CAT, if items are selected only on the psychometric properties of the items. For example, items can be selected based on the information each item provides at an examinee's given ability level

(Butterfield, 2016). There are information-based (e.g., maximum Fisher information or Kullback-Leibler information) and Bayesian item selection methods (e.g. Owen's approximate Bayesian) (van Rijn et al., 2002). The former selects the item that is most informative at a specific ability estimate. The latter selects the item based on the prior and posterior distributions of ability estimates. These methods were originally developed for dichotomous items, they have recently been applied to polytomous items (Choi & Swartz, 2009; Ho & Dodd, 2012; Penfield, 2006; van Rijn et al., 2002; Veldkamp, 2003). Maximum Fisher information (MFI) is one of the most commonly used methods (Lord, 1980). Fisher item information represents how effectively or precisely the item measures the examinee's trait level, for a given ability level. MFI selects the item maximizing the Fisher's information at a given ability estimate. After each ability estimate is calculated, the next item to be selected is the previously unused item (for that specific examinee) at the examinee's current ability level having the highest level of item information and that particular item is the item that will optimally reduce the error of the next ability estimate obtained after that item is responded and scored. (Weiss, 2011). This process results in decreasing differences in successive ability estimates as more items are administered, and it also leads standard errors of the successive ability estimates to decrease and converge during the test.

In a CAT, items can be selected depending on their psychometric properties, content balancing, and/or exposure control. Content balancing and exposure control may or may not be necessary based on the purpose of the test and use of the scores, Content balancing makes sure that the content test specifications are met and the correct proportion of items is allocated for various content areas. Kingsbury and Zara (1989) is the most commonly used content balancing procedure, comparing target

content proportions from test specifications against the actual proportions throughout the test. Exposure control ensures items are not administered to a very large number of examinees, which is important for test security (Boyd et al., 2010). Eggen and Verschoor (2006) examine whether it is possible in CATs to preserve the optimal measurement precision by administering items with a desired difficulty. The method presented in the article ensures items are chosen that have maximum information at an ability level lower or higher than the current ability estimate. An important finding of the study which is useful from the practical standpoint is that item selection targeting 60-70 % correct seems to be possible without a significant loss in precision.

Test Termination

Another important aspect of CAT design is how to terminate the exam. CATs finish once responses to a predetermined number of items have been obtained or once a conventionally acceptable level of measurement precision (e.g., $\alpha = .70$ in CTT) has been reached (Wainer et al., 2000). In CAT contexts, fixed-length and standard error termination rules are commonly applied (Gushta, 2003). Fixed-length CATs terminate once a pre-specified number of items has been administered. This termination rule has gained popularity in applied settings due to its simplicity and similarity to PBTs. However, a drawback of fixed-length termination is variation in measurement precision between individuals. Besides, fixed-length termination may decrease efficiency by terminating the test before an estimate with the minimally acceptable degree of precision is attained or alternatively by redundantly administering items that provide little or no new information about examinee ability.

Standard error-based termination (SE) is another commonly used termination rule. This rule terminates the CAT once a predetermined standard error has been

obtained (Boyd et al., 2010). An advantage of SE termination rule is that examinees are tested with similar levels of measurement precision (Choi et al., 2010). In a simulation study, Babcock and Weiss (2012) concluded that CATs terminated with a SE termination rule, as a variable-length method, discriminated highly between examinees and did not perform worse than fixed-length CATs, which is contrary to the claims made in the literature (see also Chang & Ansley, 2003; Yi et al., 2001). The findings of the study suggest that CATs terminated with the SE termination rule yielded equal measurement performance as the fixed-length versions if average test lengths are comparable. However, it is argued that compared to fixed-length CATs, variable-length CATs operated with a SE termination rule are more biased in estimates, suggesting a larger mean difference between the examinees' true ability scores and the estimated ability scores obtained by the CAT (Chang & Ansley, 2003; Yi et al., 2001). In the literature, many studies use generated data to investigate the classification performance of varying-length CATs (Spray & Reckase, 1994). However, in practice not every CAT is designed in a varying-length format. Graduate Management Admission Test and ACCUPLACER are some examples of fixed-length approaches to classification (College Board, 2003). Overall, test termination represents a key consideration in CAT design that may affect the classification of examinees.

Language Proficiency

Language proficiency refers to the general competence of an individual to communicate in the target language environment (Canale & Swain, 1980). The traditional definition of language proficiency divides linguistic skills into four: listening, speaking, reading, and writing. While this conceptualization may be practical in daily language instruction, it can lead to the over-generalization of these

basic skills (Bachman & Palmer, 1996). In other words, the tasks which are actually different in nature may be categorized under the same category. To illustrate, reading a text and trying to find out the tone of the author is very different from reading a text and trying to guess the meaning of a vocabulary item, but both would be labelled as involving reading comprehension proficiency under the above traditional definition.

As a result of the significant changes that the nature of language proficiency has undergone so far, language ability is now considered complex competence consisting of multiple inter-correlated abilities and strategies (Oller, 1979). As Zhang (2010) stated the nature of language proficiency or language ability has dramatically changed over the years. The concept of language ability as a unitary competence has gradually been replaced by the assumption that language competence is actually more complex and consists of multiple inter-correlated abilities and strategies (Bachman, 1990). For example, Zhang (2010) gives a representative example of this multi-component structure suggested by Bachman and Palmer (1996). According to this model, top tier consists of language knowledge and strategic competence. The knowledge component can be further divided into organizational knowledge and pragmatic knowledge at the second tier. Strategic competence is composed of strategies used in goal setting, assessment, and planning. At the bottom tier, organizational knowledge refers to either grammatical knowledge or textual knowledge, whereas pragmatic knowledge involves functional or sociolinguistic knowledge. According to this model, a competent speaker of a language should demonstrate the structural knowledge of a target language and also have the necessary strategies to implement that knowledge (Zhang, 2010). Likewise, Hulstijn (2011) developed a complex model of language proficiency whereby “linguistic

knowledge and the speed with which this knowledge can be processed” are the core components along with the non-linguistic components at the periphery comprised of strategic and metalinguistic competences (p. 238). According to Harsch (2014) language proficiency can be regarded as unitary or divisible based on the purpose of assessment and score reporting. Harsch (2014) also highlights the importance of reviewing research done on the nature of language proficiency outside testing contexts to have a better understanding of this competence.

In the guidelines published by Educational Testing Service (2009), factors to be considered while making decisions pertaining to testing accommodations⁶ for English Language Learners (ELL) are described. These factors which can also be seen as potential sources of construct⁷-irrelevant variance⁸ are grouped into three categories: Language Factors, Educational Background Factors and Cultural Factors. In the guidelines, language factors involve linguistic factors, proficiency level in English and in native language. Varying degrees of formal schooling in native language, formal schooling in English and exposure to standardized testing are given as educational background factors. Finally, different cultural backgrounds may place ELLs at a disadvantage in a standardized testing situation. The findings of a study conducted by Barkaoui (2014) focused on finding out whether keyboarding skill is a construct-irrelevant variable and concluded that keyboarding skill had a significant, but weak, influence on task scores. I found reading about these factors useful as they shed lights on the complexity of appropriately assessing ELLs.

⁶*Testing accommodation* refers to “any change to standardized testing conditions intended to make the test more fair and accessible for an individual or subgroup that does not change the construct being measured”

⁷*Construct* refers to “the skill or proficiency an assessment is intended to measure” (ETS, 2009, p. 12).

⁸*Construct-irrelevant variance* refers to variation in test performance originating from factors unrelated to the construct being measured.

Language Testing

Language testing theory has been dominated by two major movements. The first movement has led to definition of several scales of language proficiency on which proficiency levels are described based on the language generated by examinees evaluated against an idealized end-point: the well-educated native speaker. Scales applied to this end typically describe what examinees can do with the language and the linguistic tools used. However, according to Verhoeven and de Jong (1991), the adequacy of these scales as descriptions of developing language proficiency has been questioned. The second movement relates to Hymes (1971) sociolinguistic notion of communicative competence in models of language proficiency. These efforts have led to much more elaborated and comprehensive notion of language behavior. This movement maintains that the ability to communicate relies on factors such as personality traits and general knowledge in addition to language and the ability to use it and should therefore not be integrated in models of language proficiency.

In language teaching and testing it is common practice to divide the domain into four main skills based on mode and direction. While listening and speaking represent oral skills, reading and writing represent written skills. On the other hand, listening and reading are considered receptive skills as opposed to the productive skills: speaking and writing. Productive and receptive skills each leads to particular measurement issues. According to Verhoeven and de Jong (1991) “receptive skills are unobservable, at least in direct performance”; however, “productive skills are observable, but all that is observed in performance data is not necessarily skill related” (p. 5). Regarding language assessment procedures two opposing ideas can be distinguished. The first one involves the distinction between direct and indirect

procedures. A direct procedure refers to a procedure in which the focus of the learner is on interacting with someone in a natural way. In such a procedure, it is hypothesized that the examinee unconsciously uses the grammatical rules acquired to convey the message in spontaneous speech. On the other hand, an indirect procedure hypothesizes the learner to consciously focus on the linguistic rules required to perform the task given.

In the literature, major issues and challenges involved in testing productive and receptive skills are highlighted by different authors. Hansen and Jensen (1994) questions the validity of multiple choice to test listening. They point to the heavy cognitive load it places on listeners and suggest that “other capacities and knowledge, such as memory, reading skill and grammar, are being assessed on top of listening itself” (Hansen & Jensen, 1994, p. 259). Based on the literature, it would be fair to say that “whatever the form adopted listening skill involves extraneous processing demands that interfere with the demands of listening on its own” (Leclercq et al., 2014, p. 94). However, it is suggested that with the careful manipulation of the input, cognitive demand of the listening tests can be reduced. During this manipulation stage, the test developed must ensure that the test remains discriminatory. Washback effect of listening tests on classroom is another issue discussed in the literature. Sheerin (1987) notes that listening activities done in class turn into a form of test and listening as a skill is occasionally studied.

Oral production is one of the four basic language skills. In conversational English, the term *fluent* and *proficient* are interchangeably used. However, in language acquisition research, fluency refers to the temporal characteristics of spoken discourse such as speech rate. Segalowitz (2010, p. 31) also stated that “the prevalence of the term *fluency* in recent second language acquisition research

illustrates a scientific interest in objective, quantifiable-and therefore reliable-indicators of oral performance”. Fluency is “a performance phenomenon” (Kormos, 2006, p. 155) and fluency indicators are those that effectively uncover “how efficiently the speaker is able to mobilize and temporally integrate, in a nearly simultaneous way, the underlying processes of planning and assembling an utterance in order to perform a communicatively acceptable speech act” (Segalowitz, 2010, p. 47). For testing purposes, it appears that temporal fluency measurements are key to obtaining objective and reliable indicators of ease of processing and therefore of general spoken proficiency.

The common question posed in second language reading is whether the emphasis must be placed on second language or on reading. Both for research and testing purposes, understanding reading components are considered important to grasp how fluent readers comprehend a text. Reading processing involves important elements like word recognition, comprehension strategy use, inference making, and text model reinterpretation as a situation model (Chalhoub-Deville, 1999). There is a great deal of research available pertaining to these components and fundamental issues like the role of a reader’s prior knowledge, the relative importance of each subprocess and the extent of interaction among various subprocesses are extensively discussed.

In research on writing skill, the effectiveness of processes and interventions is usually evaluated by relating them to a measure of ‘quality of writing’, usually text quality (van Steendam et al, 2012). It is emphasized in the literature that writing is a dynamic and functional process involving various and often compensatory cognitive activities (van Steendam et al., 2012). Writing ability involves such constituents as linguistic abilities and planning skills but from writing process studies it emerges that

the whole is more than its constituent parts; that is subskills like linguistic ability do not correlate with text quality directly. In the literature, there is a considerable amount of evidence in favor of holistic evaluations of texts; however, analytical scoring is implemented in contexts where diagnostic information is required for remedial purposes.

On a different note, there are also several studies about the impact of delivery mode on writing test performance. The main aim of these studies is to look into if test-takers' performance on computer-based tests reflects the same ability measured by PBTs. In a study comparing scores assigned to paper-based and computer-based TOEFL essays, Breland et al. (2004) revealed that when test-takers were matched on English language ability, small differences were observed in effect sizes which consistently favored PBT. Wolfe and Manalo (2005), also investigating TOEFL essays, found that different test-taker demographic characteristics such second language proficiency affected the likelihood that a test-taker chooses one writing mode over the other.

Traditional language tests are often criticized for their limited construct representations. In other words, test scores from these tests may not necessarily represent a language learner's ability to cope with the language demands of tests or a curriculum. Wolf et al. (2008) highlight the serious validity concerns in language proficiency tests, especially when one considers that one of the prevalent uses of these assessment results is to make critical decisions for language students. It is important to note that language learners come from various cultural and linguistic backgrounds and have different academic backgrounds and language proficiency levels. This heterogeneity must be considered when policies and practices regarding their assessment are evaluated.

There is a body of studies available in the literature which focus on the limitations and inadequacies of language proficiency tests. The first critical limitation of traditional language proficiency assessments highlighted in these studies is that the construct of the assessment is concerned mainly with social, everyday language, and the results do not sufficiently display if the student is ready or competent enough to perform in an academic setting (Butler et al., 2007). The second limitation is lack of alignment between the language skills traditionally tested and the language demands expected in school (Stevens et al., 2000). The third limitation refers to the lack of agreement in what areas of language ability are addressed and the types of tasks used in the assessments (Zehler et al., 1994). The fourth limitation is that assessments do not involve all key language use activities (i.e., listening, reading, speaking, and writing). The fifth limitation is that the assessments are not designed to measure progress in the achievement of English proficiency in a systematic manner (Garcia et al., 2006). The final limitation refers to the issue of comparability of results from different forms of assessments. As stated by Wolf et al. (2008), this issue can be encountered when different language versions of a content area test can be implemented and investigating the content of the items and students' performance on different versions of the test is essential. Another example given by Wolf et al. (2008) is that when educational institutions are allowed to select an English language proficiency test from a list of approved commercial tests, the relationships among test scores from the different proficiency tests need to be examined so that the scores from different tests can be used appropriately.

The empirical CAT studies previously conducted in the field of language assessment have largely dwelled on its development, with some attention to validation. Most CATs concerning language testing have been designed to assess

grammar, vocabulary, and reading and very few studies focused on the assessment of listening mainly because of the complexity involved in adding a speech component to a listening test in CAT format. He and Min (2017) developed a computer adaptive EFL test (CALT) to evaluate examinees' listening and reading proficiency in English by utilizing dichotomous items and polytomous testlets. In the study, the whole process of CALT development is reported which involves item banking, determination of suitable IRT models for item calibration, item selection procedure, ability estimation method and stopping rule. The CALT in the study is a high-stakes test because it is part of the graduation requirement for non-English majors at a major Chinese university. Therefore, validity issues of the CALT are also investigated using Bachman and Palmer's (1996) assessment use argument (AUA) framework. The study provides insights into the procedures to be followed in the development of CALTs and highlighted the selection of the best IRT model for calibration of different item types.

One of the few studies done in Turkey related to English proficiency exams within IRT framework belongs to Yeğin (2003). The purpose of the study is to investigate the predictive validity of proficiency exam administered in a private university in Turkey by utilizing the 3PL IRT model ability estimates. The predictive validity was established by using the departmental English courses (DEC) passing grades of 371 students. The findings analyses showed that the 3PL model ability estimates were the best predictors for DEC passing grades were. The ability estimates obtained from sixty high information items were the second best predictor. The findings also revealed that the reading sub-test, among the three sub-tests, was found to be the best predictor of DEC passing grades.

Classification of Examinees

As would be expected in any assessment, in the assessment programs that classify students into proficiency levels for high-stakes purposes, errors and classification mistakes should be anticipated (Stone et al., 2005). Classification refers to “the act of categorizing examinees by comparing their scores on a test with given cut-offs” (Lathrop & Cheng, 2013, p. 226). Arce-Ferrer et al. (2002) give an overview of the research on sources of error in classifications accompanied with a discussion as to why the different sources have to be better understood. The sources involve sampling error, measurement error, equating error, and systematic error. The measurement error refers to the error associated with the assessment itself as well as any error associated with assigning cut scores (Linn, 2003). Zhang (2010) indicates that “by nature, classification decisions are of high risk” because misclassification may cause individuals to miss opportunities such as well-deserved educational or career development (p. 120). Classification performance especially at around the cut-off points may be of greater importance because misclassification may result in wrong decisions such as placing individuals into wrong proficiency levels.

For linear tests, classification issue has been extensively examined (Lee et al., 2008; Livingston & Lewis, 1995). However, the amount of work done so far aiming at examining classification performance of CAT, especially based on a real data set through post-hoc simulations, is relatively smaller and further research is still needed in this area. Thus, there is still a need for classification performance in CAT-based language test. Studies in classification performance of CAT is mostly based on generated data set. However, Monte-Carlo studies which are based on generated data sets may fail to represent real test-takers’ behaviour such as guessing, speediness, fatigue, etc. (Thompson & Weiss, 2011; Wang et al., 1999). The present study uses

real data set taken from a university level English language proficiency test. By using real data, optimal CAT design(s) can be defined to have higher classification performance prior to live CAT stage (Eignor & Schaeffer, 1995; Segall, 1995; Weiss, 2005). It is essential for a study on classification performance of CATs to consider that CATs, unlike linear tests, may involve different constraints such as ability estimation methods and test termination methods; thus, different constraints should also be considered when classification performance of CATs is examined.

In the literature, there are many studies examining classification performance of CATs which are specifically designed to classify individuals. These tests are called computerized adaptive classification tests (CACT). CACTs generally use varying-length test termination rule and a probabilistic approach is used to make a classification decision (Eggen & Straetmans, 2000; Gnambs, & Batinic, 2011; Rudner & Guo, 2011; Spray & Reckase, 1994; Weissman, 2007; Yang et al., 2006). CACT is a special form of CAT which seeks to classify examinees in two or more groups. In contrast to a point estimate of an individual's proficiency, the ultimate goal of CACTs is to accurately classify the examinees in various proficiency levels. Once an accurate classification decision for an examinee is reached, the testing procedure is stopped. The design of CACT necessitates numerous decisions by the test developer such as the size of the item bank, the test termination rule, and the number of classification groups.

However, not every CAT is designed in a varying-length format. Instead, there are many CATs using fixed number of items to classify individuals. Computerized adaptive tests such as ACCUPLACER and the Graduate Management Admission Test produces estimates using fixed-length approaches. Thus, classification performance of non-CACTs needs to be investigated when such

decisions are made to ensure validity and reliability of classification. Recognizing this need, the present study was conducted to examine the classification performance of fixed-length CATs on a reading subtest of a language proficiency test with respect to the PBT version of the subtest.

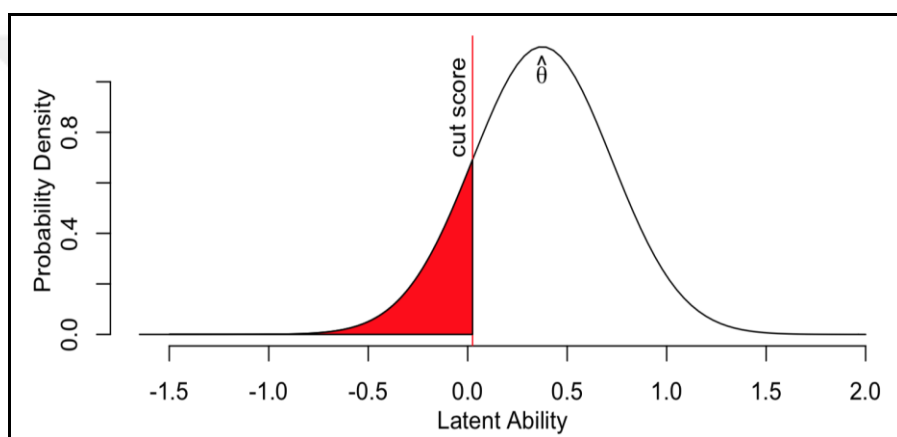
Classification Accuracy (CA) and Classification Consistency (CC) are two key indicators defining the quality of the classification decision. CC refers to the extent to which examinees are consistently classified into the same categories on two independent administrations of a test, while CA denotes the extent to which the actual classifications based on observed scores agree with the true classifications based on true scores or true classifications (Kim & Lee, 2019). Both CA and CC indices “measure the quality of the classification by quantifying the measurement error around the examinee’s ability estimate” (Lathrop, 2015, p. 2). The former is strongly related to the validity and the latter is strongly related to the reliability of the classification.

There are several approaches available in the framework of CTT and IRT to estimate classification indices either on total sum scores or latent trait estimates (Hanson & Brennan, 1990; Huynh, 1976; Lee, 2010; Livingston & Lewis, 1995; Rudner, 2005). Rudner (2001, 2005) and Lee (2010) are two widely used IRT based approaches used in many assessments. The Rudner and the Lee approach employ the same IRT model and item parameters and both construct probability distributions for each examinee. The primary difference between these two popular methods is the scale on which the classification occurs (Lathrop, 2015). Figure 5 illustrates the distribution for a single examinee produced by Rudner method. The examinee’s estimated θ is above the cut score of .0245. The area shaded in red shows the probability that the examinee is misclassified. The proportion of area under the curve

that is not red denotes CA estimate of the same examinee. His or her CC; on the other hand, refers to the proportion of the unshaded area squared plus the proportion of the red area squared, representing the probability of being classified in the same category on two independent tests. To calculate CA and CC for a group of examinees, CA and CC are estimated for each individual, and subsequently the individual CA and CC estimates are averaged to obtain the marginal CA and CC.

Figure 5

Illustration of CA and CC under Rudner's Method

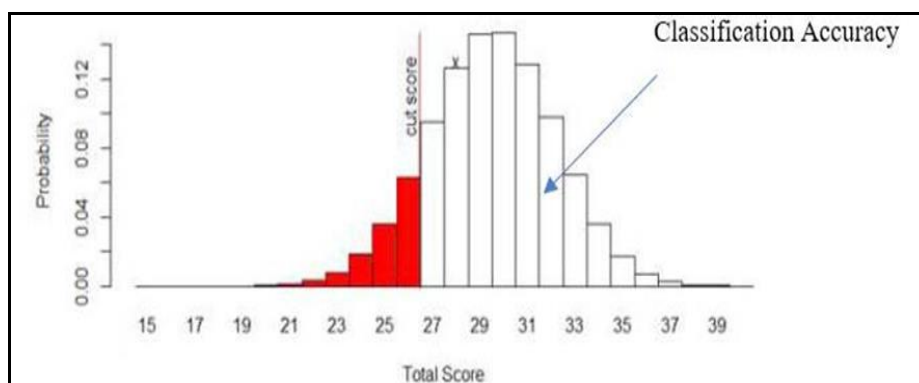


Note. From Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research & Evaluation*, 20(18), 1-5 (DOI: <https://doi.org/10.7275/43vm-p442>). Copyright 2020 by Quinn N. Lathrop.

Figure 6 shows how Lee method produces the total score distribution for the same examinee denoted in Figure 5. The examinee's total score is 28, which is higher than the cut score of 27. The proportion in red represents the probability of a misclassification. The individual CA and CC and the marginal indices are computed in the same manner as described in Figure 5.

Figure 6

Illustration of CA/CC with Lee Method



Note. From Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research & Evaluation*, 20(18), 1-5 (DOI: <https://doi.org/10.7275/43vm-p442>). Copyright 2020 by Quinn N. Lathrop.

The goal of Lee's (2010) approach is to find the distribution of all possible total scores conditional on a given θ . In the Lee approach, the classifications are estimated on the total score scale. The examinee's ability estimate and the cut score are given on the total score scale (Lathrop, 2015). Although total score scale is emphasized, the IRT model is used in the Lee approach in order to create a distribution denoting the uncertainty about the examinee's total score (Lathrop, 2015). A well-known recursive algorithm is used for this purpose (Lord & Wingersky, 1984). As a result, the distribution displays the probabilities of each total score for the examinee. Each possible total score is calculated by summing the probabilities of all possible response patterns. Table 1 gives an overview of common IRT-based approaches used to estimate CA and CC.

Table 1*Summary of IRT-Based Approaches to Estimating CA/CC*

Approach	Features	Statistical Program
Rudner (2001, 2005)	Cut-score on theta scale; mixture IRT models; need to meet IRT assumptions.	R package cacIRT (Wheadon, 2014)
Bourque et al., (2004)	Cut-score on theta scale; need to meet IRT assumptions; developed for simulation data only.	Easy to program in any programming software
Guo (2006)	Cut-score on theta scale; no normality assumptions; true item parameters or estimated item parameters close to the true parameters.	SPSS
Lee (2010)	Cut-score on observed-score scale; need to meet IRT assumptions.	R package “classify” R package “cacIRT” IRT-CLASS v2.0 for PC
Lathrop & Cheng (2013)	Cut-score on observed-score scale; does not need to meet IRT assumptions.	R package “cacIRT”

Summary

In brief, a great deal of research has been done on constructing language proficiency tests in a CAT framework and viability of CAT in language testing (Khoshsima & Toroujeni, 2017; Walczak, 2015; Wang et al., 2012). The current study contributes to the literature by examining classification performance of typical CATs on a language proficiency. Most CAT classification research has typically focused on CACTs, which are CATs specifically designed for classifying individuals into various ability groups (Eggen & Straetmans, 2000; Gnambs & Batinic, 2011; Rudner & Guo, 2011). On a different note, most studies on classification performance of CAT use generated data (Gnambs & Batinic, 2011; Thompson & Ro, 2007; Thompson, 2009), which may cause concerns about the representation of real testing conditions. The number of studies based on real data set via post-hoc

simulations, however, is relatively smaller and further research is still needed in this area. A study on classification performance of CATs also needs to consider that CATs, unlike linear tests, may involve different constraints such as test termination methods; thus, different scenarios should also be considered when classification performance of CATs is examined. In addition, the number and location of cut-off points should also be considered in such studies. The present aims to address all these aspects within its scope.



CHAPTER 3: METHOD

Introduction

This chapter outlines the methodology followed in the study. The aim of the study was (i) to investigate the comparability of CAT for a language proficiency reading subtest as an alternative to PBT, (ii) examine classification performance of CAT as compared to PBT. The study was comprised of two phases. In the first stage, comparability of results obtained from CATs under different test termination rules were analyzed with reference to PBT. To this end, a series of post-hoc simulations were conducted based on real responses to investigate the comparability of scores from CATs against the PBT version of the language proficiency reading subtest. The second stage of the study is analyzing the classification performance of CATs with various test termination rules. At this stage, the impact of the number and location of cut-off points on CAT classification performance was also examined. To this end, classification performance of CATs was compared with that of PBT using a single cut-off score at a time and multiple cut-off scores.

Research Design

The present study is a quantitative study which has two pillars. The first pillar includes post-hoc simulation studies based on actual examinee responses to a university level English language proficiency reading subtest. The ability estimations and several other statistics obtained from CAT simulations and PBT version of the test were compared.

Post-hoc simulation is a type of research design in CAT literature to find out the optimal CAT design by evaluating the effect of varying CAT algorithms under

specific conditions. The reason why this particular type of simulation was opted for in this study is that actual examinee responses were available, and they were used to run the simulations rather than generated responses as they reflect examinee characteristics better. The use of real item responses is what mainly differentiates this type of simulation from others like Monte-Carlo simulations (Thompson & Weiss, 2011). First, the item parameters are estimated by using the real examinee responses. Then, ability estimates are estimated based on the estimated item parameters. Finally, a post-hoc simulation is conducted to test every specified condition. Starting with a long test, the researcher tries to determine how short an adaptive test could be with the item bank that comprised the conventional test. In post-hoc tests, the researcher assumes that the examinees would respond in the same way to each item. In a way, post-hoc simulation provides researchers with a means of predicting how a specific CAT would perform in practice. After post-hoc simulations, ability estimates, standard errors, number of items given for each examinee were recorded for each CAT scenario.

The second pillar of the study involves a classification analysis of CAT. First, numbers of examinees who were classified into different and the same percentile were calculated. This was a descriptive analysis. Then classification indices were estimated by the method proposed by Rudner (2001, 2005). The R package cacIRT developed by Lathrop (2015) was used to calculate the classification indices. At this stage, different conditions based on two termination rules were tested (i.e., fixed-length termination and standard error termination; Gushta, 2003). To obtain a comparative analysis, the same statistics were also calculated for the real PBT. To systematically investigate the effect of various cut-off points on the ability continuum, nine cut-off values were set, creating 10 ability groups. An additional

classification analysis was made using five cut-off points to investigate the potential for the CAT to classify individuals according to the six levels of the Common European Framework of Reference for Languages (CEFR), the framework followed by the English preparatory program. Classification analyses were both conducted with one and multiple-cut-off scores at different locations on ability continuum.

Context

The response data were obtained from an English preparatory school of an English medium university in Turkey. The preparatory school is part of a private foundation university. The mission of the preparatory school, as stated on the school website, is to support the further development of students' English language and study skills in the university. It also aims to help them develop their potential as critical, analytical, and autonomous learners. As the medium of instruction is English, high quality language education is a priority for the university. The school is committed to ensuring quality testing of the English language and academic skills of its students. Students are placed into higher education programs based on nationwide university admission system.

Students registered to the university, based on their performance on the school's in-house proficiency test either start their degree programs or are placed in one of the five levels in the programme: elementary, pre-intermediate, intermediate, upper-intermediate, and pre-faculty. Students can stay in the programme for maximum two years to meet the minimum language requirements to start with their faculty programs. The examination, developed and administered in the school measure these requirements. Students need to successfully complete the pre-faculty level to sit this test. The program aims to equip students with the required level of

English and academic skills so that they could use their knowledge effectively in their departments and other aspects of life.

Participants

The data set contains 1182 student responses to 35 multiple choice items in the reading subtest of an English proficiency exam. The examinees who took the test are comprised of students who were placed into various different departments of the university at the beginning of the academic year depending on the score they obtained from the university placement system. As a result of this English proficiency test, their English level was not regarded as sufficient to study in their departments; hence, they were placed into different levels of English at the English preparatory school based on the point they obtained from the English proficiency exam. The students graduated from both public and private schools and they differ in their English language proficiency. They took the exam at the end of the academic year after receiving an approximately 10 months of instruction in English.

The number of examinees that took part in the current study was 1182 and their ages ranged from 17 to 21. The majority of the participants ($n = 984$) were receiving English instruction in the program for one or two academic years and had finished the highest level of English language instruction offered in the university English language preparatory program. The remaining participants had external preparation status and amnesty student status⁹. Upon successful completion of the university English language proficiency test, students are able to start their graduate or undergraduate degree programs in the social sciences, arts and humanities, and sciences. In this context, English is taught as a foreign language and access to the language is largely restricted to educational purposes, the internet, and popular

⁹Amnesty students are students who have reached the two-year maximum registration period in the English preparatory program but have a legal right to take the institutional proficiency exam three times after they have been dismissed.

culture. At the time of the study, 13,000 students were enrolled in the university, 3,000 of whom were registered in the English preparatory program.

Instrumentation

The data used in the study comes from the PBT version of the university English proficiency reading subtest because examinee item responses were only made available for this particular subtest by the university. The examinees' listening ability scores were also provided by the university and they were used as an external criterion.

The reading subtest used for the present study is part of an English proficiency test produced and administered in the English language preparatory school of the university. The test specifications of the school indicate that the exam tests the English language proficiency level of students with various language learning backgrounds and defines the minimum level of proficiency required for students intending to get into a graduate or undergraduate degree course at the university. The test is equivalent to 87 on the TOEFL-IBT or 6.5 on the IELTS and it is designed to assess English proficiency at the B2 level on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2002, Kantarcıoğlu, 2012). Administered three times an academic year, the institution's proficiency test is taken by around 3,000 students annually, in September, January or June. All students are required to pass this exam prior to starting in their departments. It is an in-house proficiency exam consisting of Reading, Listening, Use of Language, Writing and Speaking Skills, all reflecting academic English. In the exam, there is penalty only for the wrong answers given to the multiple-choice items. There is no cut-off score for each subtest. The examinees are expected to obtain an overall 60 percent from the exam in order to pass the exam. The exam is

administered in a single test session except for the speaking test which is scheduled at a different time. More detail about each component is given in the Table 2.

Table 2

Subtests of English Language Proficiency Test

Component	Content	Time	Contribution
Reading	3 parts consisting of a total of 6 reading texts with 35 multiple choice items. 4 wrong answers cancel one correct answer.	1 hour 20 minutes	35 points
Listening	Listening to 2 lectures in a row and taking notes after which 30 multiple choice questions are given. 4 wrong answers cancel one correct answer.	1 hour	30 points
Writing	An essay of about 350 words- a choice of two topics.	1 hour	20 points
Speaking	Engaging in a short conversation with an assessor.	Approximately 7 minutes	20 points
Language	3 parts consisting of a total of 3 tasks with 35 items. No penalty for wrong answers.	35 minutes	40 minutes

The reading test consists of three parts consisting of a total of six reading texts with 35 MC items that assess the students' ability to comprehend main ideas, read for supporting details and specific information, follow textual coherence, make inferences and guess the meaning of vocabulary from context. An item bank with 35 items may naturally be criticised from the CAT perspective. Finding items appropriate to tailor the test may be hard for the computer algorithm, resulting in lower CAT performance. However, this is an empirical study employing post-hoc simulations and was hence constrained by the research context; only items that were completed by real examinees can be included. Results are expected to provide an

insight into the potential administration of this test using CAT. Furthermore, CATs with relatively smaller number of items in their item banks have been shown to function suitably in this context. The findings reported in Sahin & Weiss (2015) demonstrated that small samples can still lead to accurate item parameter and ability estimates in CAT. In another study, CATs with as few as 20 items were shown to produce promising results with ability estimations having SE values of 0.35 (Kaya & Kalender, 2018).

Method of Data Collection

Examinees' item responses from a paper and pencil administration of an English proficiency reading subtest were obtained for the purposes of the current study. The examinee response set was provided by the relevant parties in the school upon the approval of the university. The test was comprised of 35 items (see Appendix D for the item categories and parameters) and administered to 1182 examinees. The exam is regarded as high stakes as it determines if students have an adequate level of language proficiency to cope with academic study in English. The dataset provided by the school comprises student answers in two booklet types which were merged based on the answer key. Any information regarding the student numbers and/or names were deleted from the set to protect the identity of the examinees. Furthermore, the data consisting of the examinee responses to the reading subtest of an English language proficiency test were coded dichotomously on the SPSS processor as 0 for incorrect/missing responses and 1 for correct responses. Missing rate for the items were below 3%, thus no statistical methods for imputation were employed. Cronbach's alpha of the reading subtest was .84. The correlation between reading and listening subtest scores were .77 ($p < .01$).

Methods of Data Analysis

The present study used a post-hoc (real data) simulation design. The item bank used in this study included examinee responses to all the items in a conventional paper-based reading subtest of an English language proficiency exam. Data analysis implemented involves preliminary analysis, fit-of model data analysis, post-hoc simulation analysis, comparability and classification analysis.

Preliminary Analysis

In order to investigate the impact of the missing values, two data files were formed. In one of them, 1 stood for correct responses and 0 stood for both incorrect and missing responses; in the other file, 9 was coded for the missing responses in addition to 1 for correct responses and 0 for incorrect responses. Coding the missing responses as 9 made it possible to see how the missing responses change item discrimination and difficulty values as opposed to the item discrimination and difficulty values obtained when the missing values are regarded as incorrect answers and coded as 0.

The data consisting of the examinee responses to the reading subtest of an English language proficiency test were coded dichotomously on the SPSS processor as 0 for incorrect/missing responses and 1 for correct responses. Then the descriptive statistics and frequency distribution with a normal curve were obtained to demonstrate an overall picture of the data in hand. Preliminary analysis was conducted within CTT and IRT framework so that descriptive statistics and item parameters based on PBT reading subtest can be obtained and post-hoc simulations can be carried out by using the information.

Kolmogorov-Smirnov (K-S) test was used to test for normality on the data. According to the test, normality can be assumed for this dataset, $p=.086$ at the 0.05

level of significance. Figure 7 also shows that the dataset conforms to a normal distribution with a symmetrical density curve centred around its mean.

Figure 7

Histogram of Ability Estimation by Full Length PBT

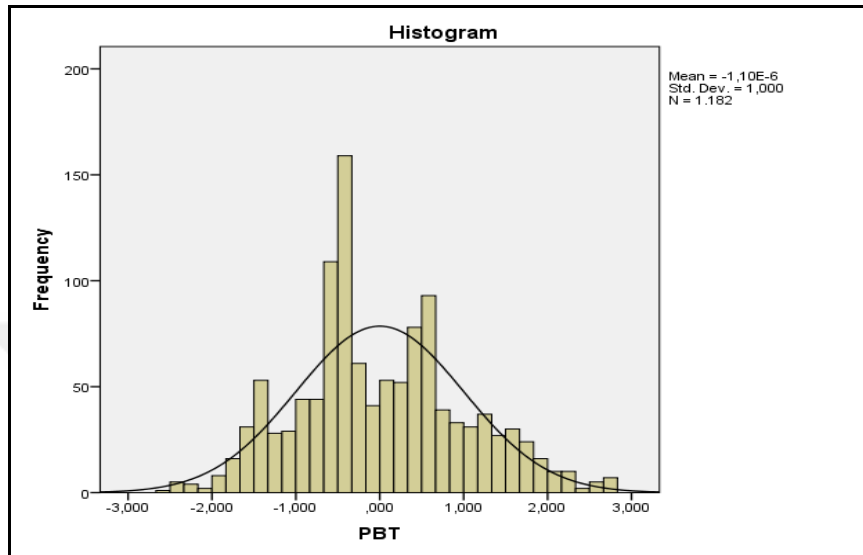


Table 3 shows descriptives of ability estimations from the full-length PBT test. It can also be inferred in the table that the dataset is normally distributed with a mean score of 0 and standard deviation of 1. The skewness value is quite close to zero, which also indicates a normal distribution and the fact that the skew value is positive shows that the data is slightly skewed to the right of the data average. The kurtosis is negative meaning that the distribution of the data is slightly platykurtic or flatter as compared with normal distribution with the same mean and standard deviation. The lowest ability estimate is -2.56 and the highest ability estimate is 2.81, indicating the existence of a wide range of ability estimates in the dataset.

Table 3*Descriptives of Ability Estimates for Full Length PBT*

Mean	SD	Skewness	Kurtosis	Min	Max
0.00	1.00	0.29	-0.23	-2.56	2.81

Fit of Model-Data Analysis

Before conducting post-hoc simulations, IRT model-data-fit was examined. The assumptions used to establish the comparability were examined. An IRT model is used to compare the CAT and PBT and therefore IRT assumptions like unidimensionality and local independence were investigated.

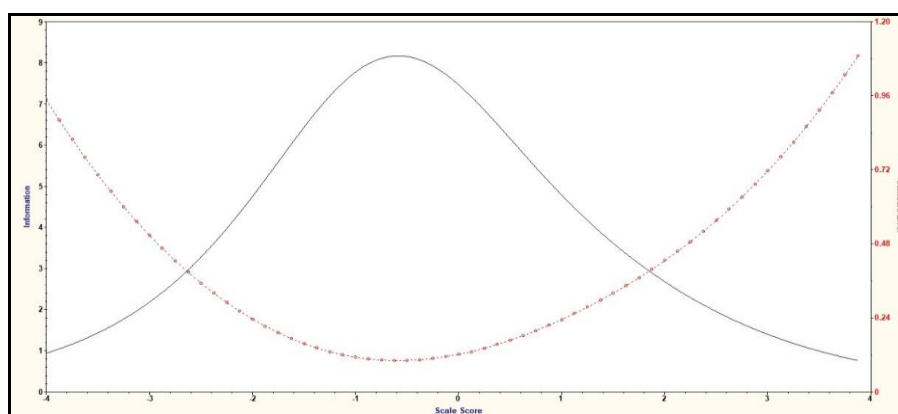
First, a factor analysis was carried out based on tetra-choric correlations using TESTFACT. χ^2 values and plot of fit data suggested that 2PL model showed the best fit for the data in hand. In the present study, BILOG-MG was used for IRT-based calibration (Zimowski et al., 1996). Results indicated that 35 items measured a unidimensional latent trait as evidenced by the ratio of the second eigenvalue to the first: 5.6 and 1.3 (Zopluoğlu & Davenport, 2017). Invariance of item parameters and ability estimates was also confirmed. Analysis of χ^2 values and plot of data fit were also examined to decide among 1PL, 2PL, and 3PL. With data to model fit, ability estimates are more reliable and the accuracy and consistency of classifications is high (Lathrop & Cheng, 2013).

Item calibration was conducted on the 35 items of the reading subtest by using BILOG-MG. Item discrimination, item difficulty and pseudo-guessing parameter were examined. To demonstrate how each item functioned, the item difficulty and discrimination indices for each item as well as for the whole test were obtained by using CTT techniques provided by BILOG-MG. In calibration phase, Bayesian expectation a posteriori (EAP) was used to estimate ability (Bock &

Aitken, 1981). The primary advantage of EAP is that they are estimating the latent trait when an examinee has only correct or incorrect responses (Desjardins & Bulut, 2018). Also, EAP method produces ability estimates with lower standard errors (Wang & Vispoel, 2005; Wang & Kolen, 2001) and is computationally more efficient than the maximum likelihood estimation method (Bock & Mislevy, 1982). An initial check of the data indicated the existence of monotonic response patterns, which further justified the selection of EAP as the ability estimation method. Mean (standard deviation) of item discrimination parameters was 0.895 (0.316) whereas item difficulty parameters had a mean of -0.247 (0.932). Minimum and maximum ability estimates were -2.057 and 2.81 ($M=0$, $SD=1$). Standard error of PBT had a mean (standard deviation) of 0.343 (0.250). Forty-eight percent of the individuals had a standard error of 0.40 or lower, whereas 90.4% had 0.50 or lower. Figure 8 shows the item information curve of the PBT.

Figure 8

Item Information and Posterior SD Curves for PBT



Note: Straight line shows item information and dotted line shows posterior standard.

Post-Hoc Simulations

A real-data simulation (i.e., a post-hoc simulation) of CAT is an important step in CAT research. Such a simulation helps CAT developers to assess important

features of the CAT such as item selection and test termination rules before administering the test. The main reason why real data simulation is because use of generated data (Monte-Carlo simulations) may not be reflecting examinee's psychometric characteristics (Wang et al., 1999). In a post-hoc simulation, the matrix of item responses in a post-hoc simulation is comprised of real responses from a conventional test of a certain length. When actually executing a CAT, the test implementer would not have control over test aspects such as the degree of misfit and the parameter distributions, so post-hoc simulation would allow researchers a means of predicting how a particular CAT would perform in practice. A post-hoc or a real-data simulation proceeds as follows (IACAT, 2016):

1. IRT parameters are estimated for the items in the test
using appropriate software based on the available item response data.
2. Theta for each examinee is estimated by maximum likelihood or Bayesian methods by using those parameter estimates.
3. Based on the item responses available for each examinee from the conventional test administration, CAT with maximum likelihood (or Bayesian) theta estimation is applied to adaptively estimate theta for all the examinees.
4. CAT is terminated when a pre-specified termination criterion has been reached.
5. CAT theta estimates are compared against the conventional test theta estimates.
6. Adaptive test lengths are determined to obtain maximum similarity between the CAT theta estimates and those of the conventional test, with a minimum number of CAT items.

It is essential for a study on comparability and classification performance of CATs to consider various constraints such as ability estimation methods and test termination methods. For a regular CAT, different constraints can be defined such as an ability estimation method and an item selection method. The present study focuses only on test termination rule as a constraint. Other constraints like ability estimation and item selection methods are not within the scope of the analysis done in the study. The test termination rule plays an important role in determining the length and difficulty of the CAT and is most susceptible to criticism from various stakeholders such as the community, parents or students. When a large-scale high stakes CAT defines a termination rule based on precision of ability estimates, test takers answer different numbers of items and this may be a point of public criticism of the testing program. In tests involving the same number of items, the fact that the test involves different items for each test taker may still be objectionable to examinees, families and other parties unless efforts are made to raise levels of CAT literacy among stakeholder groups (O'Loughlin, 2013).

Simulations were set to be terminated based on fixed-length and SE termination rules. In the current study, 10 different CATs (five fixed-item CATs and five se-based CATs) were defined to simulate for each examinee. As for SE-based CATs, five standard error rates were selected (i.e., below 0.50, 0.40, 0.30, 0.20, 0.10) corresponding to different alpha values from .75 to .99 and labelled as SE05, SE04, SE03, SE02, and SE01 (Weiss, 2011). This variable-length rule terminates the CAT once the SE of the ability estimate falls below a pre-specified threshold representing the desired level of precision, which is often taken to be 0.3, corresponding to a reliability of 0.9 (Nunnally & Bernstein, 1995). Apart from SE-based CATs, CATs terminated after FL10, FL15, FL20, FL25 and FL30 items were also simulated.

Terminating a CAT after 10 or 15 items may raise questions about the content coverage or the content validity of the test. However, these two test termination scenarios were also used in the study because it may prove informative for future research. For this study, all CAT simulations were conducted by a software developed by Kalender (2011).

During the post-hoc simulation, the initial ability estimation was set to 0 for all examinees and CAT software randomly selected an item with moderate difficulty indexes (-0.05 and +0.05). Next, the software checked the student's responses from the PBT data set and depending on the student's response, the software estimated an ability and selected another item based on the estimated ability level using maximum information function. Bayesian EAP procedure was used to estimate individuals' ability in all simulations. The CAT procedure goes on selecting items with maximum information based on the examinee's response to the selected item. Fisher's maximum information was employed to select items to minimize the standard error associated with examinee ability estimation and maximize test information (Matteucci & Veldkamp, 2013; Thissen & Mislevy, 2000). Once the test termination criterion was met, the simulation stopped. Thus, the mean of the ability estimations was used in the analyses after the simulation was replicated 100 times for each examinee and with each stopping rule to identify any possible bias that the starting procedure might cause to the ability estimations. At the end of simulations, an ability score, its standard error, number of items given over 100 replications were recorded into a file for each individual.

Comparability Analysis

Due to defensibility and accountability concerns as well as professional testing standards, whenever paper- and computer-based assessments of the same

content are administered, empirical evidence showing comparability of test scores obtained in the two administration modes is required. Otherwise, in the absence of such evidence, CAT cannot be proposed as a viable alternative to the PBT version of an exam. In this study, only the psychometric category of the framework developed by Wang and Kolen (2001) is used. In the light of this framework, comparability of CATs was evaluated with reference to the PBT was evaluated depending on various stopping rules as a CAT administration condition. To this end, item reduction, correlation, ability estimates and standard errors were obtained and investigated for each CAT test termination rule simulated, as suggested by Wang and Kolen (2001).

Classification

The way CATs are designed and delivered render such tests relatively more challenging in nature for examinees compared to PBTs. Considering this, prior to computing CA and CC estimates through Rudner approach, a percentile rank analysis of CATs was done on SPSS without using a specific approach chiefly for descriptive analysis. Examinee's ability estimates were divided into 10 deciles on SPSS based on examinees' theta values in CAT and PBT. The decile of each examinee's estimated ability value in PBT is compared against that in different CAT scenarios in order to investigate how the ability percentiles of the examinees vary across CATs with reference to PBT. The number of examinees placed into the same or different percentiles was calculated. Theta values of each examinee in PBT is checked against those in every CAT scenario to see if there is a significant percentile change for each examinee. This initial percentile rank analysis proved helpful to visualize the level of change in each examinee's rank across various CAT scenario with reference to the one in PBT.

Following this, Rudner approach, relying heavily on IRT, was used to estimate CA and CC across various ability distributions and cut score locations. Rudner approach is chosen mainly because in this approach, the cut score is placed on the θ (theta) metric rather than the observed-raw score scale. To implement this method, a freely available statistical R package “cacIRT” developed by Lathrop (2015) was used to compute CA and CC indices. The Rudner approach functions with an examinee’s latent ability estimate along with a cut score on the same latent scale. Rudner approach was preferred because the study is based on computer adaptive tests where “examinees respond to different items and possibly different numbers of items” (Lathrop, 2015, p. 4). The assumption behind the Rudner approach is that “the ability estimate and its standard error form a normal distribution” (Lathrop, 2015, p. 2). The R package cacIRT provides implementation of Rudner approach in a unified framework and code syntax. As part of this analysis, the effect of using different cut-off points on the ability continuum was also explored. To this end, nine cut-off values were defined from 10th to 90th percentiles increasing in increments of 10, which created 10 ability groups using normal distribution. The cut-off points on the ability continuum corresponding to percentile scores of each decile were -1.35, -0.73, -0.51, -0.42, -0.17, 0.25, 0.51, 0.81, 1.42. Another classification analysis was done on six categories with five cut-off points to investigate the potential for the CAT to classify individuals according to the six levels of CEFR followed by the English preparatory program. Five cut-off points were set using 16.6, 33.3, 50, 66.6 and 83.3 as percentile ranks. These cut-off points were -0.92, -0.46, -0.12, 0.40 and 0.91. CA and CC values were computed both for a single cut-off score at a time and a simultaneous analysis of all cut-off scores.

CHAPTER 4: RESULTS

Introduction

This chapter presents the results obtained from multiple rounds of analyses under two main sections. In the first section, results regarding comparability of scores estimated by PBT and CAT are presented. In the second section, results of the classification analysis are given accompanied by the relevant figures and tables.

Comparability Analysis

Ability estimates from CAT and PBT were compared with respect to differences in the ability estimations, item reduction rate, correlations, and the reliabilities. Table 4 below shows mean and standard deviations of estimated ability values in the various CAT scenarios over 100 replications. Mean and standard deviation of PBT-based ability estimates were set to 0 and 1, respectively, for reference. The table denotes small negative means values for all ability estimations, which reveals that there is only a negligible discrepancy across CAT and PBT versions of the reading subtest in terms of estimations. According to the results presented in Table 4, it can be concluded that different CAT strategies delivered tests with similar difficult levels, as evidenced by similar ability scores.

Table 4

Descriptives of Ability Estimates by CATs Averaged over 100 Replications

CAT	<i>M (SD)</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Skewness</i>	<i>Kurtosis</i>
FL10	-0.02 (0.88)	-2.347	2.093	0.036	-0.407
FL15	-0.02 (0.92)	-2.427	2.409	0.128	-0.243
FL20	-0.02 (0.94)	-2.566	2.570	0.176	-0.226

Table 4 (cont'd)*Descriptives of Ability Estimates by CATs Averaged over 100 Replications*

CAT	<i>M (SD)</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Skewness</i>	<i>Kurtosis</i>
FL25	-0.02 (0.95)	-2.440	2.675	0.214	-0.229
FL30	-0.02 (0.95)	-2.513	2.574	0.228	-0.265
SE01	-0.02 (0.95)	-2.432	2.594	0.265	-0.299
SE02	-0.02 (0.95)	-2.432	2.594	0.265	-0.299
SE03	-0.02 (0.95)	-2.432	2.594	0.266	-0.296
SE04	-0.01 (0.94)	-2.432	2.594	0.278	-0.241
SE05	-0.02 (0.92)	-2.331	2.594	0.284	-0.104

To establish the extent to which ability estimates vary according to PBT, descriptive statistics of ability estimates were calculated and a series of paired-samples t-tests were performed to statistically compare the mean ability estimates by CATs and PBT. Table 5 shows mean and standard deviations of estimated ability values in various CAT scenarios (PBT mean was set to 0.00) averaged over 100 replications and the results of the t-tests. The table denotes smaller standard deviations and negative means for all CATs, indicating the ability distributions with CATs are narrower, though not considerably. The t-test results indicated statistically significant mean differences between FI15, FI20, FI25, FI30, SE01, SE02, SE03, SE04 and the PBT version, however, effect sizes of these differences were small as estimated by Cohen's (1988) *d*, indicating that ability levels estimated by CATs and PBT are similar.

Table 5*Means (SDs) of the Ability Estimates by CATs Averaged over 100 Replications*

CAT	M (SD)	t	p	d	CAT	M (SD)	t	p	d
FL10	-0.02 (0.88)	1.51	.13	-	SE05	-0.02(0.92)	1.95	.05	-
FL15	-0.02 (0.92)	2.44	.01	0.06	SE04	-0.01(0.94)	2.73	.00	0.07
FL20	-0.02 (0.94)	2.55	.01	0.07	SE03	-0.02(0.95)	9.19	.00	0.22
FL25	-0.02 (0.95)	3.20	.00	0.09	SE02	-0.02(0.95)	9.96	.00	0.30
FL30	-0.02 (0.95)	5.08	.00	0.14	SE01	-0.02(0.95)	9.96	.00	0.30

Table 6 illustrates the number of items used in CAT simulations by using fixed-SE CAT scenarios. Expectedly, the number of items used increases with low SE values. It is clearly observed in Table 6 that no reduction in the number of items is obtained with 0.20 and 0.10 SE threshold values. All examinees were given all the items. However, a relatively significant reduction in the number of items (as compared to 35 items in PBT) can be achieved with 0.50 and 0.40 SE threshold values, which correspond to $\alpha \geq 0.75$ and 0.84, respectively (Cronbach, 1990). Overall, CATs with SE thresholds of 0.50 and 0.40 produced promising results compared to the other test termination scenarios.

Table 6*Descriptives of the Number of Items used in SE CATs*

CAT	M (SD)	Reduction in Number of Items (%)	Minimum Number of Items	Median
SE05	14.64 (7.27)	58.17	7	13.00
SE04	26.63 (7.77)	23.91	14	27.00
SE03	34.10 (2.30)	2.57	34	35.00

Table 6 (cont'd)*Descriptives of the Number of Items used in SE CATs*

CAT	M (SD)	Reduction in Number of Items (%)	Minimum Number of Items	Median
SE02	35.00 (0.00)	0.00	35	35.00
SE01	35.00 (0.00)	0.00	35	35.00

Table 7 shows the number of students who were given all of the 35 items by CAT with respect to different SE thresholds. With the CATs terminated at SE values of 0.20 and 0.10, all 35 items were administered to all 1182 examinees. No reduction in the number of items was obtained for any examinees. If the CAT is terminated at 0.30 SE threshold, 1010 examinees (85%) were administered all 35 items. The rest of the examinees were administered less than 35 items. CAT seems to yield a smaller number of items at SE threshold values of 0.40 and 0.50. Particularly in the CAT terminated with 0.50 SE threshold value, the examinees were administered less than 35 items.

Table 7*Descriptives of Examinees Administered 35 Items for SE CATs*

CATs	<u>n</u> ~~N~~	% of Examinees
<0.50	62	5.25
<0.40	427	36.13
<0.30	1010	85.44
<0.20	1182	100
<0.10	1182	100
PBT	1182	100

The magnitudes of 0.40 and 0.50 SE threshold values for fixed test length CAT formats were further examined for tests containing 10, 15, 20, 25 and 30 items since these two SE threshold levels indicate reasonable reliability levels. The results of the analysis are presented in Table 8 and 9 for 0.40 and 0.50 SE values, respectively. Increasing the number of test items led SE estimations to decrease. For tests containing 10 items, none of the ability estimations for the examinees fell below an SE of 0.40. With 15 items, there is a small percentage of examinees with ability estimations below an SE value of 0.40. The mean SE values obtained for 10 and 15 items are relatively higher, which shows that the tests with these CAT formats estimated the individuals' abilities with a lower reliability, compared to the tests with 20, 25 and 30 items. As can be seen in Table 9, with fixed-length CAT format, the mean standard error values are at 0.40 or higher for tests with 15, 20, 25 and 30 items. Overall, the simulation results clearly show that a significantly lower number of items can be administered if the CAT version of the test is terminated at 0.50 SE threshold value. As Table 6 shows, an average of 14 items is administered which equates to 60% of reduction in the number of items compared to the PBT version of the reading test.

Table 8

Descriptives of Ability Estimations at ≤ 0.40 SE for FL CATs

FL CATs	%**	Mean	Median	Mode	Minimum	Maximum
10 items	0	-	-	-	-	-
15 items	9.47	-0.52	-0.44	-0.40	-1.36	-0.33
20 items	26.48	-0.66	-0.48	-0.45*	-1.47	-0.26
25 items	36.29	-0.47	-0.45	-0.48*	-1.50	0.51
30 items	45.09	-0.36	-0.41	-0.56*	-1.52	0.59
PBT	49.74	-0.33	-0.45	-0.53	-1.64	0.68

* Multiple modes exist. The smallest value is shown.

** The percentage of examinees' standard error of their ability estimates is equal to or less than 0.40.

Table 9

Descriptives of Ability Estimations at ≤ 0.50 SE for FL CATs

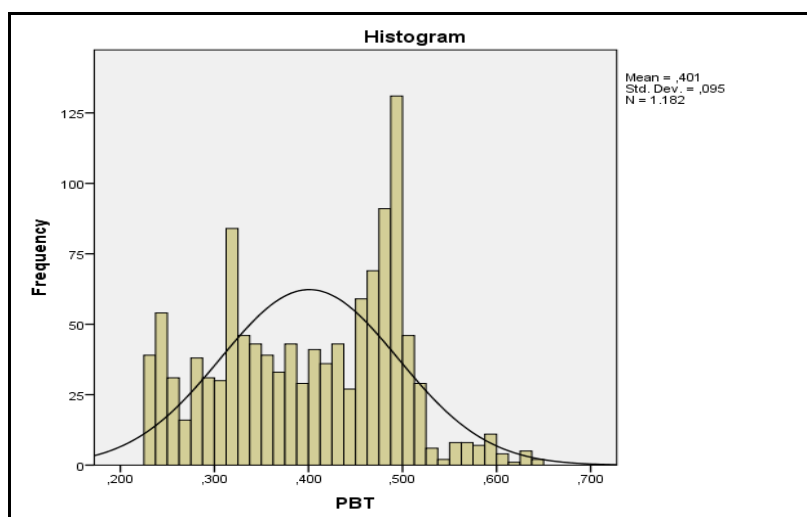
FL CATs	%**	Mean	Median	Mode	Minimum	Maximum
10 items	40.43	-0.70	-0.64	-0.53	-1.56	-0.13
15 items	71.82	-0.37	-0.34	-0.25*	-1.75	0.65
20 items	80.62	-0.34	-0.34	-0.45*	-2.36	0.81
25 items	78.84	-0.17	-0.24	-0.48*	-2.44	1.42
30 items	76.14	-0.13	-0.23	-0.56*	-2.51	1.57
PBT	89.17	-0.16	-0.32	-0.53	-2.56	1.69

* Multiple modes exist. The smallest value is shown.

Figure 9 shows how 0.50 SE based CAT worked. It can be inferred from the histogram that a significant part of the ability estimations (69%) from the full-length PBT fall below 0.50. Based on the histogram, it can be inferred that CAT produced similar levels of error for ability estimations.

Figure 9

Histogram of SE Values for PBT Ability Estimations



For fixed-length CAT simulations terminated when 10, 15, 20, 25 and 30 items were given, the mean SE values were analyzed. Table 10 below presents the results of the analysis. The negative relationship between number of items given and SE can be seen. The mean values for the CATs terminated with 20, 25 and 30 items are around 0.40. The CAT terminated with 10 items as a rigorous test termination rule led to a SE value which is higher than 0.50.

Table 10

Mean SE Values for FL CATs

Fixed-Length CATs	Mean of SE Values
10 items	0.53
15 items	0.47
20 items	0.44
25 items	0.42
30 items	0.40

In the present study, the ability estimates calculated based on various test termination criteria were correlated with the ability estimates obtained from the full PBT test. Table 11 gives Pearson correlations of ability estimates between CAT and PBT tests across all conditions. All the correlation values given in the tables were significant at the Alpha level of .01. With every termination criterion employed, CAT produced highly correlated results with the ability estimations obtained from the PBT.

Table 11

Pearson Correlations between PBT and CAT Ability Estimations

SE threshold	<i>r</i>	Fixed-Item	<i>r</i>
<0.50	.952	30 items	.994

Table 11 (cont'd)*Pearson Correlations between PBT and CAT Ability Estimations*

SE threshold	<i>r</i>	Fixed-Item	<i>r</i>
<0.40	.991	25 items	.985
<0.30	.999	20 items	.974
<0.20	1.000	15 items	.957
<0.10	1.000	10 items	.927

As an external criterion, the examinees' raw scores on the listening subtest were utilized to validate the ability estimations from CATs obtained from using different test termination rules. Table 12 indicates the correlations between CAT ability estimations with the listening subtest scores across all the test termination criteria. Overall, all the correlations are higher than .70. All the correlation values given in the tables were significant at the alpha level of .01. However, the correlations between the ability estimations based on the examinees' listening scores from PBT and those from CATs terminated with standard error threshold rule are slightly higher than the correlations based on fixed-length CATs.

Table 12

Pearson Correlations between CAT Ability Estimations and Raw Scores of Listening Subtest

SE threshold	<i>r</i>	Fixed-item	<i>r</i>
<0.50	.767	30 items	.766
<0.40	.767	25 items	.761
<0.30	.767	20 items	.753
<0.20	.767	15 items	.741
<0.10	.767	10 items	.719

Classification Performance of CAT

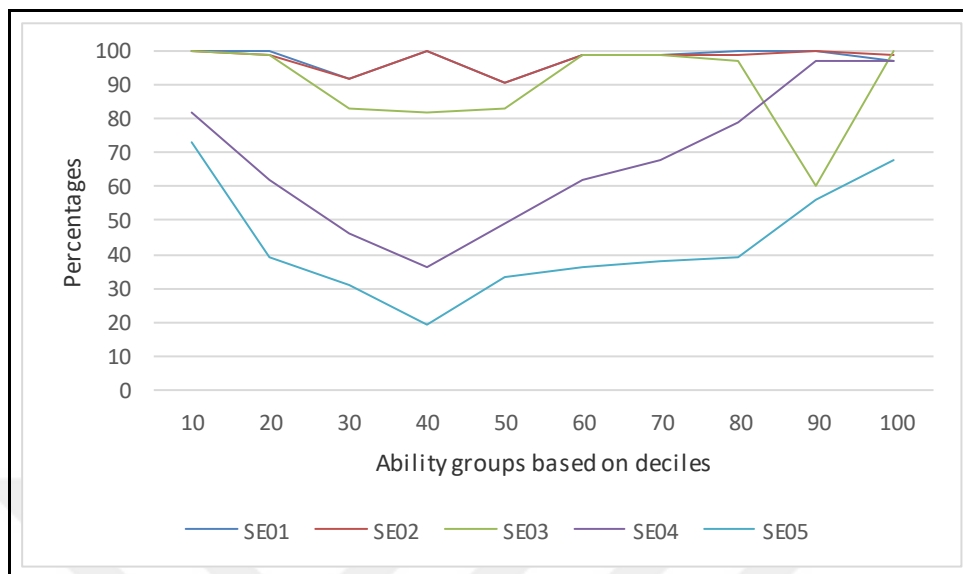
Percentile Rank Analysis in CAT Results with Reference to PBT

Examinees were divided into ten ability groups based on deciles to check how the locations of examinees on ability continuum change in PBT and across different CAT scenarios. Ability estimation of each examinee in PBT is checked against those in every CAT scenario to see if there is a significant percentile change for each examinee. The results of the analysis are shared below.

Figure 10 displays the rates of examinees (in percentages) placed in the same or one level above/below with respect to their original rank in PBT. There is a clear relationship between the number of examinees classified into the same group and SE values. For lower SE values, more examinees were classified into the groups. For example, stricter CATs terminated at lower SE values like 0.10 and 0.20 produced better classification rates. By such CATs almost all examinees were put into the same ability groups with reference to their rank in PBT. However, as shown below, these CATs do not have any item reduction. Starting from the SE value of 0.30, a reduction in the number of items can be observed. With 0.30 SE value, the classification rates seem higher. Another trend is observed for the same CAT scenario in terms of location of groups on ability continuum. CATs terminated at 0.40 and 0.50 SE values classified less examinees into the same ability group compared with middle ability groups. The same trend can also be observed in Figure 10. With SE 0.40, classification rates drop, especially around the middle percentiles. Most of the examinees in the highest and lowest percentiles are classified in the same percentile in every CAT. For the CAT terminated with SE 0.50 whereby the number of items administered decreases, the classification rates significantly drop, except for the highest and lowest percentiles.

Figure 10

Percentages of Examinees Correctly Classified into the Ability Group by SE CATs



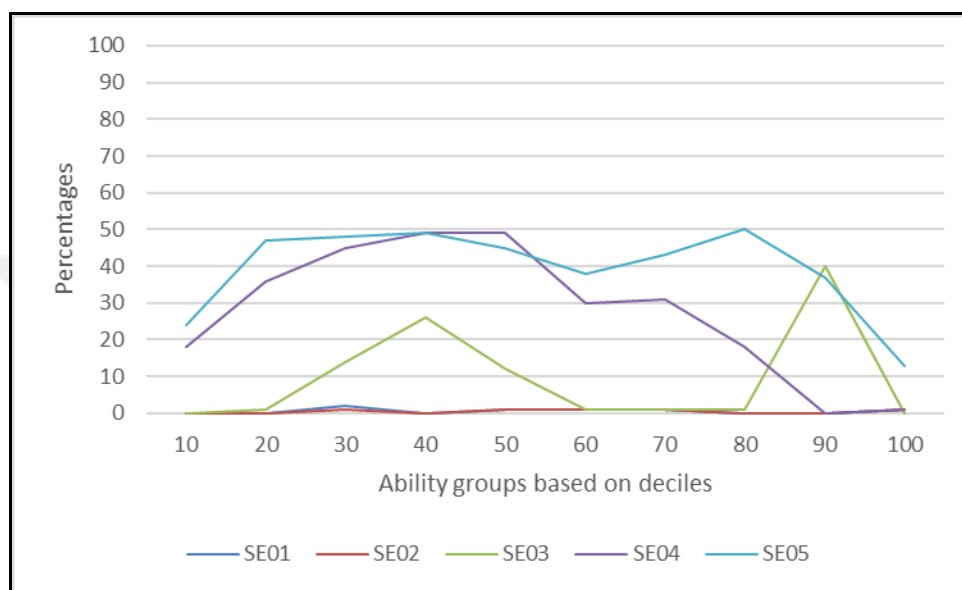
To further investigate the classification rate of each CAT with different stopping rules, the percentages of individuals who were not classified in the same way but fell into only one class higher or lower than their original percentile in their PBT scores were also checked (see Figure 11). The majority of the examinees who did not fall into their original percentile in different CATs were classified only one percentile higher or lower, which helps us infer that CAT can classify students precisely. For example, for the CAT terminated at 0.50 standard error threshold, in the 40th percentile, the percentages of the examinees classified exactly the same as PBT is 36 and the percentages of those classified either one decile up or down is 49. Overall, it would be fair to say that the CAT actually performs well in classifying examinees with respect to the examinees' percentiles in PBT.

When rates of examinees who were put into a group one above or below by different CAT scenarios were examined in Figure 11, it was seen that the number of examinees who were ranked in the same percentile by CATs as their original rank in PBT tends to decline as the SE values increase (see Figure 11 and see Appendix A

for the tabulated data). It can also be observed in the same table and figure that, for each CAT scenario, the examinees were placed only one ability group higher or lower, if not in their original rank in the PBT.

Figure 11

Percentages of Examinees Classified into One Group Above or Below by SE CATs



The same classification comparison was also made between PBT and fixed-length CATs. As can be seen in Figure 12 and 13, the rate of examinees classified into the same percentile rank or one rank above/low is less in comparison with CATs operating with SE test stopping rule. However, as in the case of CATs terminated with SE threshold stopping rule, for fixed-length CATs, the percentages of the examinees who were ranked in the same rank and one rank above or below adds up to 100 or a value near 100 in every percentile, showing that the examinees were placed only one ability group higher or lower, if not in their original rank in the PBT.

The results show that classification performance of fixed-length CATs varies in line with the number of items (Figure 12 and 13). As the number of items administered in CATs increase, the classification of the examinees become more precise. For example, in the third group from the top, the rate of examinees falling

into exactly the same percentile as the one in PBT systematically increases from 32% in the 10-item CAT up to 80% in the 30-item CAT. In addition to the number of items, the location of the percentile rank seems to affect the classification performance of CATs. For every fixed-length CAT, the examinees in the middle percentiles are classified less precisely. To illustrate, for the 30-item CAT, it can be observed that the rate of the examinees classified into the same rank as the one in PBT significantly drops from 96% in the 10th percentile to 66% in the 60th percentile. The classification rate starts to rise again starting from the 60th percentile. A similar trend can be observed in other fixed-length CATs. Overall, classification rates for all the percentiles provide relatively more stable and higher results in 25-item and 30-item CATs with reference to PBT. In the middle percentiles, the rate of examinees classified into the same rank as the one in PBT tends to decline; however, it is also observed that even then most examinees do not fall into any further than one percentile higher or lower than their original rank.

Figure 12

Percentages of Examinees Correctly Classified into the Same Group by FL CATs

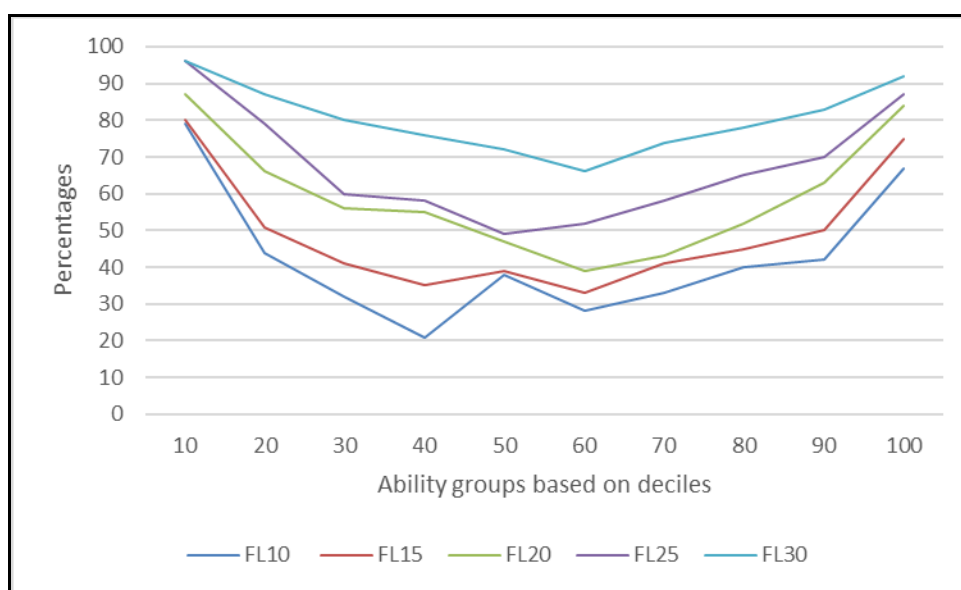
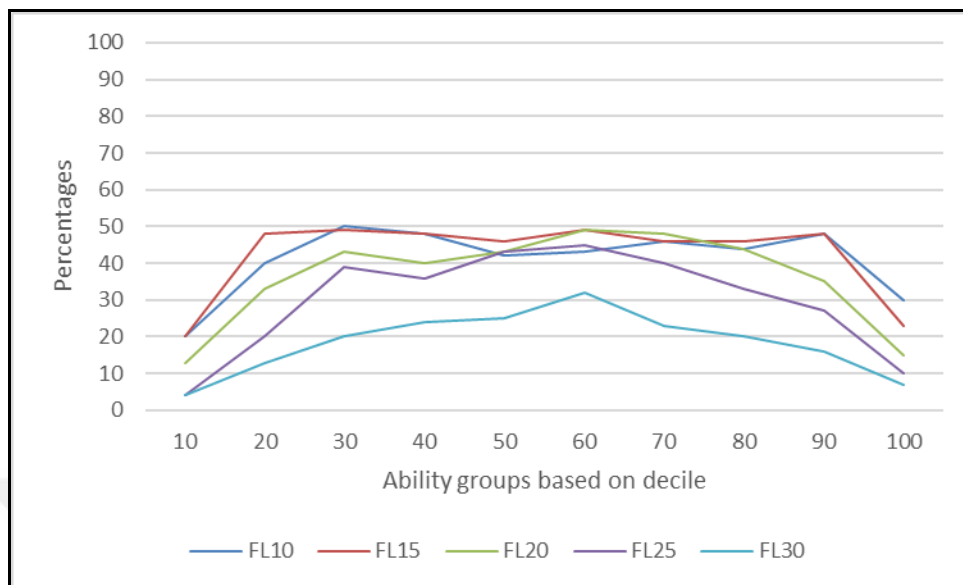


Figure 13

Percentages of Examinees Classified into One Group above or below by FL CATs



In brief, percentile rank analysis gave a general idea about whether examinees are classified by CAT into the same categories as PBT over replications or to what extent the examinees' rank vary with reference to their original rank on PBT. Overall, the classification performance of all the CATs terminated with fixed-length and standard error threshold rules was similar. With reference to PBT, most examinees are classified into the same group in the low and high ability groups but examinees tend to move to either one percentile up or down around the middle ability levels. There is not a significant difference between SE and FL CATs in terms of the mobility of the examinees.

For every fixed-length CAT scenario, the vast majority of the examinees were classified either in their original percentile rank in the PBT or only one percentile higher/lower. It was observed that classification performance was almost perfect especially for the examinees whose ability estimates fall within the first (10th) or the last (90th) percentile rank. A similar trend can be observed in CATs terminated with 20, 25 and 30 items. Classification performance slightly fell for CATs with 10

and 15 items. For example, for CATs with 15 items, 77 examinees out of 1182 were not classified in the same percentile as PBT or one percentile higher/lower. They were classified in a different percentile. Overall, the results show that higher number of items lead more examinees to fall into the same percentile as their PBT classification. For example, with 80th percentile, the percentages of examinees classified into the same percentile are 40, 45, 52, 65 and 78 for FL CATs with 10, 15, 20, 25 and 30 items respectively.

As for CATs terminated with various SE threshold values, it was observed that classification performance was almost perfect with reference to PBT, especially for the examinees whose ability estimates fall within the first (10th) or the last (90th) percentile rank. However, more examinees were classified into one group up or down rather than their original PBT percentile through the middle ability groups. For example, with CATs terminated with 0.40 standard error threshold value, while in the 10th percentile 82 percent of the examinees were classified into the same decile, in the 40th percentile, 36 percent of them were classified into the same decile. On a different note, it was also observed that the percentage of examinees who fell into the same decile as their PBT classification significantly increased as the SE values fell. With the 40th percentile, for example, the relevant percentages are 19, 36, 82, 100 and 100 for SE05, SE04, SE03, SE02 and SE01 respectively.

Following this initial percentile analysis, which showed a descriptive picture of CAT classification performance, a more systematic analysis was done based on Rudner's method.

Estimating Classification Accuracy (CA) and Classification Consistency (CC)

Following the percentile rank analysis, classification performance of CATs was analyzed by computing CA and CC values with Rudner method (2001, 2005).

The R package *cacIRT* developed by Lathrop (2015) was used to calculate CA and CC, both of which range from 0 to 1, with higher values indicating higher classification performance (e.g., a CA value of .76 indicates that there is a 76% chance the individual has been appropriately classified and the same value for CC means the probability of being classified into the same group across two administrations). The Rudner approach was preferred to other approaches because in CAT context examinees respond to different items and different numbers of items and therefore it would not be meaningful to classify examinees based on their total score (Lathrop, 2015). To examine the effect of using different cut-off points on the ability continuum, nine cut-off values were defined from 10th to 90th percentiles increasing in increments of 10, which created 10 ability groups using normal distribution (-1.35, -0.73, -0.51, -0.42, -0.17, 0.25, 0.51, 0.81, 1.42). Classification analysis was done by using a single cut-off score for each CAT.

Figures 14a-d below present the results of the CA and CC estimates computed. To avoid complexity, Figures 14a-d show CA values for CATs with fixed-item and SE-threshold separately. CA values for PBT (demonstrated by the line with the legend PBT) provide a reference point for comparison. The same applies to Figures 15a-d. The CA and CC values were obtained using only one cut-off score at a specific percentile. In other words, each dot in each line indicates a CA or CC value assuming there is only a specific cut-off score.

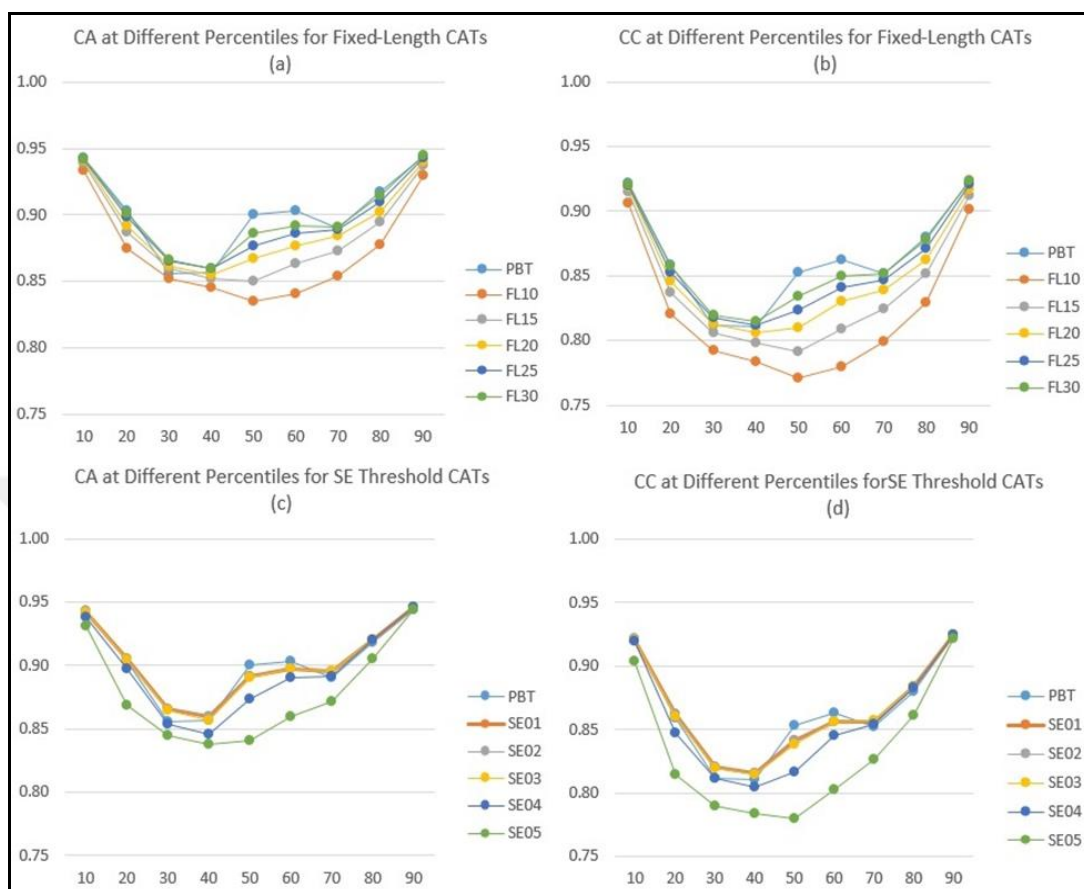
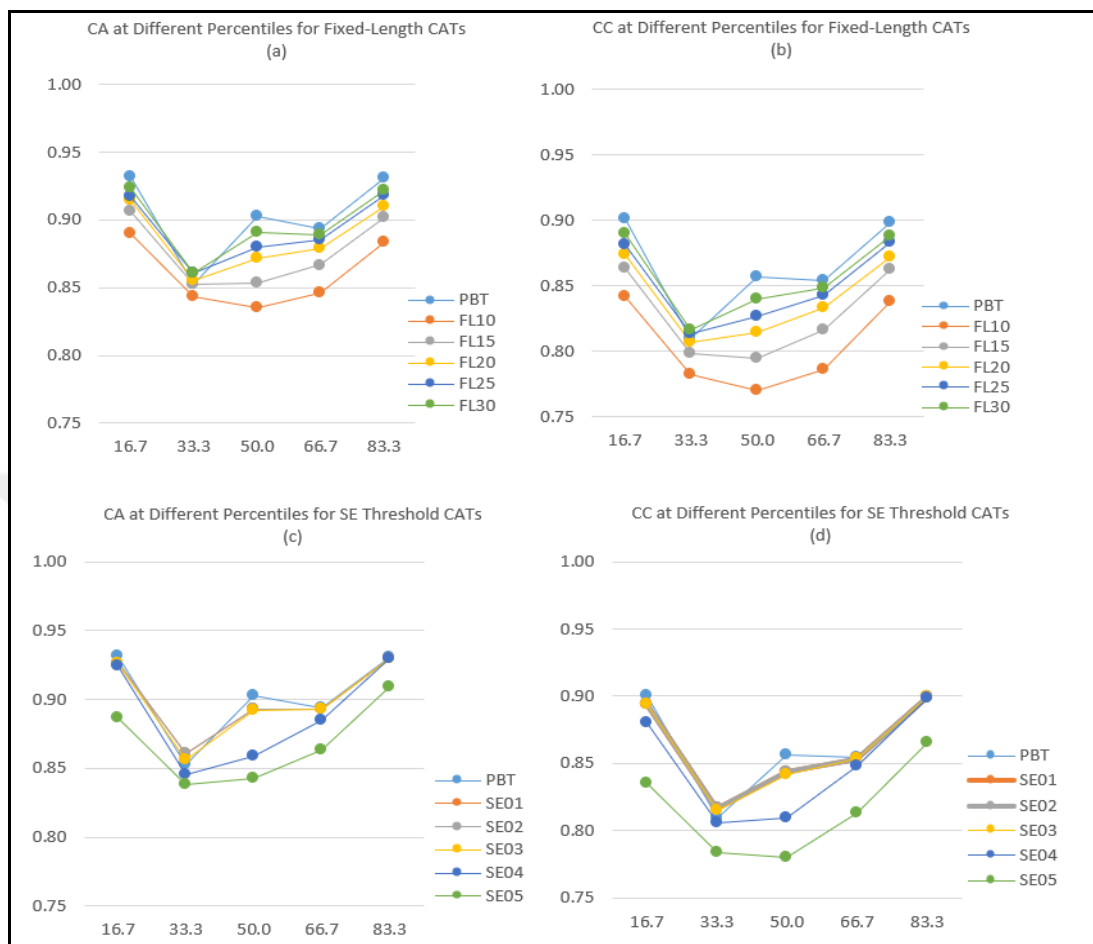
Figure 14a-d*CA/CC Values with 9 Cut-Off Points for Different CATs*

Figure 14a-d shows that both CA and CC estimates for all CATs with different termination rules are higher than 75 percent. This indicates that at least 75% of individuals can correctly be classified into the ‘true’ groups and this can repeatedly be observed across different and parallel administration. It is also observed that for all CATs CA estimates are slightly higher than CC estimates. When the CA values are further examined based on the scenarios under two different test stopping rule, it is seen that greater variability exists in CA estimates obtained from fixed-length CATs. The same trend can be observed in CC estimates as well. Because no item reduction was achieved in the CATs terminated with 0.10 and 0.20 standard error threshold values, CA and CC estimates are very close to the PBT

estimates. Declining CA and CC values can be observed for the CATs working with 0.30, 0.40 and 0.50 standard error threshold values as the number of items administered also decreases. Another very clear trend is that the CA and CC estimates show an upward trend at the higher and lower tails of ability estimates. It seems that the higher the ability estimate is, the higher CA and CC estimates are for all CATs regardless of the CAT scenario applied. The same trend can be observed for the lower ability groups. Especially for the groups in first and last percentile, CA and CC values are estimated to be above 90%.

Besides, CATs with fixed-items produce differing CA and CC values around middle ability range (see Figure 14a and 14c) especially at the 50th, 60th and 70th percentiles, whereas SE-based CATs have relatively more coherent CA and CC estimates (see Figure 14b and 14d). Both for CA and CC estimates, CATs have values at lower and higher ability levels similar to the ones in PBT; however, around the middle ability levels, PBT has higher values for all CAT scenarios. It is also observed that PBT outperforms CAT regardless of the test termination rule used, particularly in the 50th and 60th percentiles. In the lower and higher ability groups the gap between CAT and PBT classification values significantly narrows. CAT scenarios with lower number of items or higher SE values (FI10 and SE05) lead to relatively lower CA and CC values around middle ability ranges. It would be fair to say that the classification performance of CATs with 25 and 30 items is above 85% for the average ability range and 95% for the high ability estimates, which is very similar to the classification performance in the PBT version of the subtest. Similarly, CATs with SE04 had CA and CC values around 85% at the middle ability range and 95% in the tails of the ability range.

To examine the potential for the CAT to classify individuals according to the six levels of CEFR followed by the English preparatory program, an additional classification analysis was done with five cut-off points. Five cut-off points were set using 16.6, 33.3, 50, 66.6 and 83.3 as percentile ranks. These cut-off points were -0.92, -0.46, -0.12, 0.40 and 0.91. CA and CC values were calculated both for one cut-off score at a time as well as a simultaneous use of all cut-off scores. Examinees were grouped into 6 ability strata by using 5 cut-off scores. By doing so, more heterogeneous ability groups were created. Using the similar approach with 10 ability groups, CA and CC values were estimated using one cut-off point each time. The results of the classification analysis with 5 cut-off scores, as shown in Figures 15a-d, yielded a similar pattern to the one above. With five cut-off points, a similar pattern was observed around middle ability range for both CA and CC values. CATs showed different classification performance around that range. Like classification with nine cut-off points, CATs with fewer items or higher SE values showed relatively lower classification performance. However, unlike classification analysis with nine cut-off points, CA and CC values are different around the high and the low ability levels. Generally, classification performance decreased in the tails. For example, CATs with 10 and 15 items or CATs with SE of 0.50 and 0.40 showed lower performance both in CA and CC. Both for nine and five cut-off points, CATs performed equally or lower than PBT in classification.

Figure 15a-d*CA/CC Rates with 5 Cut-off Points*

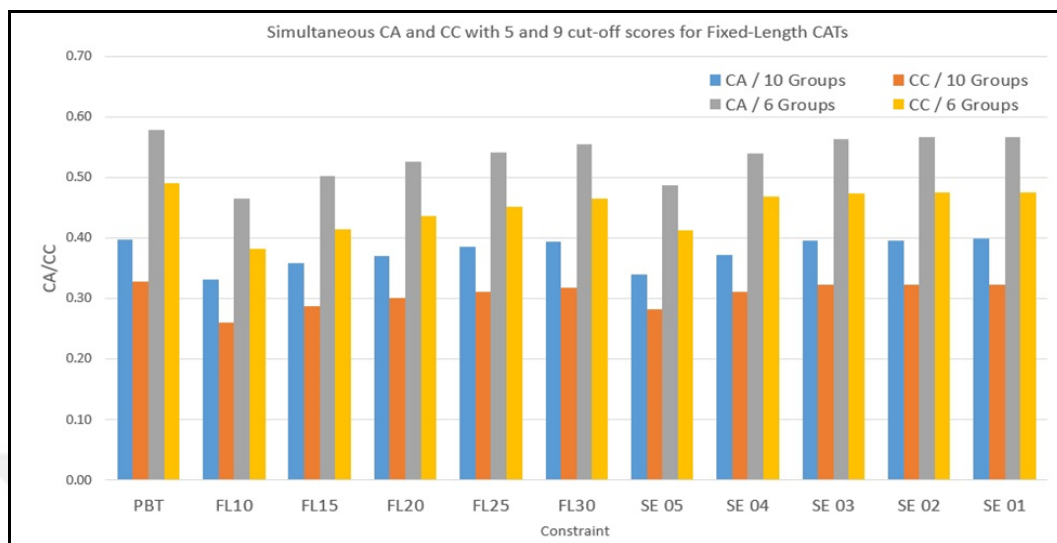
With five cut-off points, a similar pattern was observed around middle ability range for both CA and CC values. CATs showed different classification performance around that range. Like classification with nine cut-off points, CATs with less number of items or higher SE values showed relatively lower classification performance. However, unlike classification analysis with nine cut-off points, CA and CC values are different around the high and the low ability levels. Generally, classification performance decreased in the tails. For example, CATs with 10 and 15 items or CATs with SE of 0.50 and 0.40 showed lower performance both in CA and CC. as compared to the classification with 9 cut-off points, there is a decrease at the tails of ability continuum in terms of classification performance as evidenced by

lower indices. This is especially true for FL10 and SE05 CATs because group homogeneity in the tails reduce.

However, instead of using one cut-off score at a time, multiple cut-off points were generally used in classification. Using the same schemes above, another two sets of classification was conducted using the nine and five cut-off scores together. Figure 16 shows CA and CC values for each CAT scenario when 9 and 5 cut-off points were operated simultaneously and it can be clearly seen that CA and CC scores significantly decreased compared to classifications with a single cut-off score. Classifications with a single cut-off scores had CA and CC values ranging from 75% to 95%, whereas simultaneous classification produced CA and CC values between 25% and 60%. The results of the analysis also show that for all CAT scenarios, CA and CC values tend to rise significantly when the number of cut-off points fall. The classification with 9 cut-off scores produced lower CA and CC values than the one with 5 cut-off scores. In general, different CAT scenarios produced similar CA and CC values for a given number of cut-off points. The figure also shows that with fixed-item CATs, higher number of items leads to higher CA and CC values. However, in all cut-off point arrangements, CA and CC values for PBT are the highest although differences are minor. With the CATs terminated at SE01 and SE02, CA and CC values increase and they are the same as those PBT produces.

Figure 16

CA/CC Values for Simultaneous Classification with 5 and 9 Cut-Off Points



In brief, with a single cut-off score, CAT-based classification performed very well but not for every ability level. Classification worked better for high and low ability groups in all CAT scenarios, but not as good as PBT. Besides, classification performance dropped significantly when a simultaneous classification analysis was carried out with multiple cut-off points. CA and CC values significantly decreased compared to classifications with a single cut-off score. Classifications with a single cut-off score recorded CA and CC values ranging from 75% to 95%, whereas simultaneous classification produced CA and CC values between 25% and 60%. The classification with nine cut-off scores produced lower CA and CC values than with five cut-off scores. Number of cut-off points were found to be as a significant predictor of classification performance. Lower cut-off points yielded higher values of classification indices.

Summary

Key findings obtained from the series of analyses can be outlined as follows:

Comparability

- Smaller standard deviations and negative means for all CATs suggest that the CATs were more difficult, though not significantly.
- The paired sample t-tests conducted to statistically compare the mean differences between ability estimates by CATs and PBT indicated statistically significant differences, however, effect sizes of these differences were small indicating that ability levels estimated by CATs and PBT are similar.
- No reduction in the number of items was achieved using 0.20 and 0.10 SE thresholds, however a relatively large reduction can be obtained with 0.50 and 0.40 SE thresholds.
- The mean SE values for the CATs terminated with 20, 25 and 30 items revolve around 0.40 and higher number of test items leads to lower SE estimations.
- Reliability of estimates by CAT was comparable with reference to PBT as evidenced by the correlations of ability estimates between CATs and PBT.
- CATs with termination criteria produced statistically significant correlations with reference to ability estimates obtained in the PBT.

Classification

Percentile Rank Analysis

- Percentile rank analysis showed that with reference to PBT, most examinees are classified into the same group in the low and high ability groups but examinees tend to move to either one percentile up or down around the middle ability levels.

- There is not a significant difference between SE and FL CATs in terms of the mobility of the examinees.
- For every fixed-length CAT scenario, the vast majority of the examinees were classified either in their original percentile rank in the PBT or only one percentile higher/lower.
- It was observed that classification performance was almost perfect with reference to PBT in CATs terminated with 20, 25 and 30 items.
- Higher number of items leads more examinees to fall into the same percentile as their PBT classification.
- As for CATs terminated with various SE threshold values, classification performance was almost perfect with reference to PBT, especially for the examinees whose ability estimates fall within the first (10th) or the last (90th) percentile rank. However, more examinees were classified into one group up or down rather than their original PBT percentile through the middle ability groups.
- It was also observed that the percentage of examinees who fell into the same decile as their PBT classification significantly increased as the SE values fell.

Classification Analysis with Rudner Method

- CA and CC estimates show an upward trend at the higher and lower tails of ability estimates when nine cut-off points are employed. With five cut-off points, a similar trend was observed around middle ability range for both CA and CC values.
- CATs with fewer items or higher SE values showed relatively lower classification performance, for both 9 and 5 cut-off scores.

- With the use of five cut-off points, CA and CC values decreased in the tails unlike classification analysis with nine cut-off points.
- For all CAT scenarios, CA and CC values rise significantly when the number of cut-off points drop.
- For all cut-off point arrangements, CA and CC values for PBT are higher than those for CATs though the differences are minor.
- Classification of high and low ability groups was particularly enhanced by using CAT.
- Classification performance dropped significantly when a simultaneous classification analysis was conducted with multiple cut-off points. With a single cut-off score, CAT-based classification performed quite well but not for every ability strata.

CHAPTER 5: DISCUSSION

Introduction

The discussion of the research findings reported in this chapter are divided into parts corresponding to the research questions. At the beginning of each part, the relevant research question is restated and an overview and interpretation of the findings which give insights into the related question are provided. Following this, these findings are discussed with reference to the literature review. At the end of the chapter, limitations of the study are given and suggestions for implications and future research are made.

Overview of the Study

When a computerized adaptive test (CAT) is considered to be used in a testing programme along with a PBT or as an alternative to it, the comparability between these test versions needs to be researched to ensure that the scores are not biased in favor of a particular test procedure. To this end, the present study first examined comparability between CATs with different testing conditions and PBT versions of a language proficiency reading subtest by considering various comparability criteria. The comparability analysis constitutes the first stage of the study and addresses the first research question with its sub-questions.

The second stage of this dissertation focuses on classification performance of a typical CAT, which is not specifically designed for classification. CAT has potential to improve classification decisions in placement testing in the context of English medium university admissions. However, if CAT is to be considered a realistic alternative, it is important to compare the classification performance of various CATs with reference to PBT as this comparison may prove useful in the

design process of CAT. Despite the increasingly prevalent use of CATs to classify examinees, two classification indices, CA and CC, have mostly been discussed with reference to linear tests (Lee, 2010; Lee et al., 2008), and these indices are relatively unexplored in the CAT literature. In addition, CAT-based classification has not been sufficiently examined with data real examinee responses (Cheng & Morgan, 2012) and the effect of systematic examination of location and number of cut-off points on classification performance of CATs is currently unclear.

The problems mentioned above are particularly important for non-CACTs, CATs not designed specifically for classification purposes. A CAT targeting classification generally uses varying test length for examinees. However, institutions may prefer to offer fixed-length CATs. This is one of the decisions that language programs face is about whether to use varying-length CATs or fixed-length CATs. This is an important decision since it does involve several parties such as institution itself, examinees and families. Classification performance may differ significantly across the two approaches (Cheng & Morgan, 2012). Another decision language programs must make is whether to include single or multiple cut off points in the CAT. Binary decisions like mastery or nonmastery are common in testing (Eggen, 2009). However, location of the cut-off points used for a binary decision is also a concern. Classifying individuals into more than two groups for placement purposes is a commonly expected function of locally developed language tests (Dimova et al., 2020). The present study explored these gaps in the literature by examining the classification performance of CAT based on real data consisting of examinee responses to a locally developed and administered language test. All analyses of classification were performed by comparing the results to PBT.

In brief, the present study sought to compare ability estimates and other parameters estimated by CATs. The other purpose of the study was to investigate classification performance of CATs under different testing conditions. To this end, an initial percentile rank analysis was carried out to obtain a general picture regarding the classification performance of CATs across various ability groups. With this percentile rank analysis, the number of examinees correctly classified into the same or a comparable ability strata was examined. Following this, a more systematic analysis was conducted using Rudner method to estimate CA and CC values for every CAT scenario. The results were compared considering test termination rules and the number of cut-off points. For both purposes, values for PBT provided a reference point for comparison.

Discussion of Major Findings

The first question along with its sub-questions asked about the extent to which fixed-length and SE-based CAT administrations of English proficiency reading subtest produces comparable ability scores to PBT in terms of test difficulty, items administered, standard errors, and the correlations between ability estimates obtained from CATs and PBT. Results of the comparability analysis showed that there was not a significant difference between CATs and PBT in terms of test difficulty. The results indicated that there was a negligible difference between all CATs and PBT, which is expected. A CAT is inherently more difficult than the conventional PBT version as the items are tailored to the ability level of the examinee during the exam every time a response is given and measurement efficiency is the core principle of a CAT (Ling et al., 2017; Samsudin, et al., 2020). Thus, examinees are less likely to receive items that fall below their ability level. Correlations between the PBT and all CAT ability estimates were high (Pearson $r=$

.93-1.00). Overall, the simulation results clearly showed that a significantly lower number of items can be administered if the CAT version of the test is terminated at 0.50 and 0.40 SE threshold values or at 15, 20, and 25 items under fixed length test termination rule.

The simulation results clearly also showed that there was no reduction in the number of items using 0.20 and 0.10 SE thresholds. This is mainly because such low SE thresholds are relatively stringent as test termination rules and the number of items used in the simulations is small since the item bank was not originally developed for a CAT, which always necessitates a large item pool (Weiss, 2004). On the other hand, a relatively large reduction can be achieved with 0.50 and 0.40 SE thresholds, which correspond to Cronbach's $\alpha \geq 0.75$ and 0.84, respectively (Cronbach, 1990). When using these two reliability levels, it was observed that only a small number of examinees were given all 35 items, 62 (5.25%) and 427 (36.13%) out of 1182 for CATs respectively. In the study, as expected, it was observed that reduction in item number administered is closely linked to the amount of standard error, which is consistent with the literature in the field of CAT.

One of the conclusions drawn by the research done so far is that when SE threshold is opted for as a test termination rule, by increasing SE threshold, the researchers are able to do ability estimation with less number of items, which is supported by the literature (Bulut & Kan, 2012; Kreiter et al., 1999; Kezer 2013). On a related note, as also stated by Özbaşı (2014) and by Weiss (2011), it would be useful to reiterate these findings because the item bank used in this study was not originally developed for CAT and the number of items was limited, there may not be an item available in the bank for each ability estimate in the sample. Babcock and Weiss (2012) highlighted that an item bank may not have sufficient number of items

to satisfy a rigorous termination criterion, which is the case in this study. CATs with 15, 20, and 25 items also yield promising results with acceptable standard errors. With each termination criterion used, CAT led to highly correlated and statistically significant correlations ($p < .001$) with ability estimates attained in the PBT.

In general, no significant differences were observed between PBT and CATs terminated with SE threshold and fixed-item stopping rules. Regardless of the test termination method, CATs produced comparable ability estimates with its PBT counterpart. This is a key finding as it provides a freedom of choosing either of the test termination methods for the test developers as both result in equal quality in measurement, which is especially helpful from a practical standpoint (Bobcock & Weiss, 2012).

Obviously, it is important to note that the decision regarding which test termination rule to utilize has to be taken also after considering the social and even legal consequences of test termination alternatives, particularly in countries like Turkey, where the stakeholders demand transparency in matters of educational assessment and communication between test developers and stakeholder groups is highly expected, which is also considered by O'Sullivan (2016) as a central component of validation. Considering the high-stakes nature of the language proficiency exams at university preparatory programmes and criticism that is likely to be made by students and families of CAT implementation (different items, different number of items, etc.), the results highlight the need for test developers to be flexible in their approach toward test termination. Asking the same number of items to every examinee may create a feeling of equality. To conclude, the core principles of CAT (i.e., a reduction in the number of items administered, individual reliability estimates, and comparable results with PBT) were observed in the present

study prior to classification analyses and the results of the comparability analysis showed that CAT can be considered as an alternative to PBT.

The second research question and sub-questions of the study asked about the classification performance of fixed-length and SE-based CAT administrations of English proficiency reading subtest with reference to PBT. To this end, a percentile rank analysis was conducted in response to the first sub-question. The percentages of individuals who were classified into the same percentile by CATs and PBT were calculated. An additional analysis was done to see if the percentages of examinees classified into one percentile were higher or lower than their original PBT rank. Ability estimates of each examinee in PBT were checked against those in every CAT scenario to see if there is a significant percentile change for each examinee.

The results revealed that the percentages of examinees who fell into a percentile group different from their percentile was quite low. All CATs correctly classified examinees into the same or the neighbouring strata. This initial percentile rank analysis was useful to see the level of change in each examinee's rank across various CAT scenario with reference to the one in PBT. The results showed that CATs were able to place the majority of the examinees into the same percentile as the one in PBT or in one percentile higher or lower.

Following the percentile rank analysis, which is descriptive in nature, accuracy and consistency of the classification decisions were analysed by estimating CA and CC through Rudner's method to address the second sub-question under the second research question. The results obtained from classification analysis demonstrate that CA and CC values are lower than those estimated for what would be expected in CAT regardless of the test termination method, similar results were also highlighted by Mizumoto et al. (2017). Considering this study used only the

items in the original PBT reading subtest, the fact that PBT worked better in CATs was somewhat expected.

Cheng and Morgan (2012) and Lathrop and Cheng (2013) reported higher levels of CA and CC values in CAT in their studies. However, their studies involved larger item banks and content balancing methods. Considering that no such method was employed in this study, the findings are nonetheless promising. The results indicate that there is potential to introduce CAT in this context with a larger item pool.

As the analyses showed, not all scenarios provided equal CA and CC values. Classification was better for CAT scenarios with FL30 and SE01 around the middle ability range whereas CATs with FL10 and SE05 provided lower classification performance by almost 10% at the same range. CA values are around .80 while CC values go down to .75 for the simulations. Overall, CATs cannot outperform PBT, which is to be expected as the simulation was necessarily based on 35 items. However, the results indicate that there is potential to introduce CAT in this context with a larger item pool. Furthermore, the analysis indicated that not all scenarios provided equally defensible CA and CC values. Classification was better for CAT scenarios with FL30 and SE01 around the middle ability range whereas loose CATs with FL10 and SE05 provided lower classification performance almost 10% at the same range.

Overall, the results provide a paradox. CATs with a strict termination classify the examinees better but they require a larger item bank. Less stringent CATs like FL30 also perform well in terms of classification but only with a small reduction in the number of items administered. Related to this, it would be fair to say that a

balance needs to be created between addressing the core principles of CAT and classification.

The third sub-question under the second research question asked about the difference between CA and CC values on the PBT and CAT versions of the test with different test termination rules when one cut-off point was used at a time. As Figures 14a-d and 15a-d in Chapter 4 show, CA and CC values are lower than what would be expected in CAT regardless of the test termination method (i.e., CAT is expected to increase measurement precision over PBT; see Mizumoto et al., 2019 and Gyllstad et al., 2021 for a discussion of CAT expectations in language testing). CA values are around 80% while CC values go down to 75% for the simulations. PBT-based CA and CC values are either higher or equal to the CA and CC values for both termination methods.

When a single cut-off score was utilized for a binary decision, such as pass or fail, classifications with one cut-off score recorded CA and CC values ranging from 75% to 95% for both fixed-length and SE-based CATs. For classifications with one cut-off point, the graphs are comprised of lines with a U-shape curve, suggesting that the classification performance for low and high ability groups is higher than for mid-ability groups due to the absence of disproportionately difficult or easy items. Besides, at the tails, CA and CC estimations are above 75%, and reach around 90% in every fixed-item and SE-based CAT. A U-shaped trend was observed in each analysis, suggesting that CAT achieves higher CA and CC for low and high ability groups than for the middle ability groups. These findings of the study are in parallel with those reported in the literature. Lathrop and Cheng (2013) found a similar U-shaped trend for various conditions (i.e., different test lengths, IRT models, and cut

score locations) and there was a significant fall in CA performance around the center of the ability distribution.

In relation to the fourth sub-question under the second research question, the effect of operating multiple cut-off points at the same time was also examined for all CAT scenarios. The results demonstrated the classification performance of CATs with five cut-off points was better than those with nine cut-off points. A reduction in the number of cut-off points from nine to five, increased CA and CC values even at the tails of the ability range, which is possibly due to the location of cut-off points along the ability continuum. Decreasing the number of points from nine to five also changes the locations and the distances between them, creating more heterogeneous groups, supporting the findings reported by this research. This finding suggests that a lower number of cut-off points creates groups with less homogeneity in the ability continuum. Increasing the cut-off points and decision categories naturally results in a higher amount of classification error because multiple categorizations and comparisons are made (Cheng & Morgan, 2012). The CATs with five cut-off points consistently perform almost 10% better than all the other CATs (see Figures 4a-d). CA estimations with five cut-off and nine cut-off points are around 40% and 60%, respectively, which is in line with the findings reported by Cheng and Morgan (2012). In their study, the number of decision categories was one of the factors that significantly varied classification performance of CAT in which simulated item responses were utilized. The current study confirms their findings using real responses to a real test.

In brief, the findings of the current study demonstrate that the decision to use multiple cut-off points simultaneously needs to be taken cautiously as CA and CC drop with the increasing number of decision categories used. Given the number of

examinees in high and low ability groups is often a lot lower than the number in the middle ability group, as in a typical normal distribution, classification performance around the middle level actually carries more importance since it influences more examinees. Even if the decision to be taken about the examinees is binary (pass or fail) and the cut-off point is located around the middle ability group, the sensitivity of the score becomes more critical for all the stakeholders. Thus, classifying individuals around the middle ability range requires a lot of attention. Apart from this, the findings of the study also show that mismatch between the mean difficulty level of items and the ability levels of examinees needs to be taken into consideration while designing CATs for classification purposes. Examinees may exhibit aberrant response behavior such as blind guessing when item difficulty and ability levels are not aligned, which may, in turn, cause problems with estimation of ability levels with sufficient precision and classification of examinees correctly. This problem can be overcome by including items in the item bank that provide more information around the cut-off points at the middle ability range and making sure that there is a better match between item parameters and examinee ability (Berger et al., 2019).

Results of the study proved promising in terms of various psychometric aspects including classification performance demonstrating that the language proficiency reading subtest can be administered using CAT. However, it is important to note that the present study looked into CAT from psychometric perspective. If the decision is taken to convert to CAT, developing a legally defensible CAT will require more than psychometric expertise. The practical sides should be researched. The primary consideration is if the program has the capacity to develop extensive item banks to cater for a wide range of ability levels, which was particularly emphasized by Babcock & Weiss (2012) and Weiss (2011). The other

important consideration is if the program has or can afford the necessary resources (e.g., computers, laboratories and CAT delivery engine). Whether the program has the psychometric expertise and whether converting the test to CAT is likely to bring the expected reduction in test length are other points to consider before making the decision to transform an assessment program from paper-based fixed form tests to CAT. On a related note, it is worth indicating that every testing program has distinct issues that must be identified and isolated to be investigated empirically through the use of simulation research because adequate research and documentation both before and after the development of CAT will lay the foundation for an efficient and legally defensible test. The choices made at every stage must be backed up with empirical evidence to obtain accurate and reliable examinee scores.

Although it was originally stated as a research question, in response to the feedback received during the peer review process of the manuscript based on this dissertation, a content analysis was also carried out to examine the construct coverage of the CATs in the simulations following classification analysis. Item distributions across the simulations were investigated and the 25 most frequently used items were recorded (see Appendix B). Items were placed into five categories (Chikalanga, 1992): propositional inference (14 items), paraphrase of supporting detail (14 items), pragmatic inference (4 items), meaning of unknown lexis in context (2 items), and textual cohesion (1 item) independently by two teachers in the English preparatory program with 100% agreement. As can be seen in Table C, five of the simulations were involved in this analysis because there was no reduction in SE01, SE02, and the error of ability estimates were higher in FL10 (0.53), FL15 (0.48), and FL20 (0.44). The results show that the item categories were unequally represented across the simulations. For instance, the item assessing textual cohesion only features

in SE03. Although this finding may seem to indicate a significant reduction in construct coverage, it results from the disproportionate number of items associated with each category. That is, rather than evidence of systematic bias in item selection between the reading subdomains by the item selection algorithm, the unequal distribution is a result of the limited items identified as testing categories D and E.

Implications for Practice

The findings of this study have key implications in the context of educational measurement. A great deal of research concerning classification in CAT aims to classify examinees into binary categories like pass or fail, the method and the findings of this study are meaningful in contexts where multiple cut-off scores are used (Eggen & Straetmans, 2000; Spray, 1993; Weissman, 2004). The major finding of the study is that classification performance of CAT varies depending on the number of decision categories and cut-off locations. The results clearly demonstrate that CA and CC values are higher when limited number of decisions are to be made based on the test results. This finding supports the conclusions drawn by Cheng and Morgan (2012) that CA and CC fall if classification decisions are made on the basis of more sophisticated categorization.

The present study has also implications specific to the field of language testing. As there was no substantial difference between test termination rules, test developers have more freedom in the design process of CATs. Not only fixed-length CATs but also SE-based CATs may prove useful for language testing, which is a crucial piece of information if CAT is introduced into the current context. Test termination decisions have social and even legal implications, particularly in Turkey in which the public expect transparency in matters of educational testing and especially high-stakes decisions based on language tests are often contested in

courtrooms. The finding that the CAT termination method does have a small effect on the classification of individuals has crucial implications for public debates surrounding assessment and this should be communicated to various stakeholder groups (Chalhoub-Deville & O'Sullivan, 2020). Considering the high-stakes involved in language proficiency exams at university preparatory schools and criticism that may be received from students and families about CAT process (factors such as different items and different number of items), the results allow test developers to be flexible in their approach toward test termination.

The content analysis demonstrated that various reading abilities were measured irrespective of the test stopping rule. However, it was also observed that certain abilities were better represented than others, suggesting that the practical consideration of content balancing procedures in CAT design would be useful to ensure the items from the different reading subdomains are administered. Using CAT in a reading comprehension test including items that aim to measure a variety of reading processes may lead to an uneven representation of certain items between the examinees. This variation raises a key question: if two statistically equivalent tests vary substantially in terms of content, will stakeholders be convinced that the test measures the same reading construct? As indicated in the literature review, CAT classification research mainly concentrates on the potential to maximize measurement precision. The effect of various test termination rules on construct coverage has been neglected. Reading tests feature a range of items targeting different cognitive processes (Khalifa & Weir, 2009). In academic reading tests, the construct involves lower-level processes such as scan and skim mechanisms as well as more complicated cognitive processes such as inference (Bax, 2013). Sufficient construct coverage of a test is important. Otherwise, the test score may not inform

valid decisions about test candidates' academic reading ability (Weir et al., 2009). On a different note, item selection algorithms increase measurement precision potentially at the cost of the test construct as certain cognitive operations may be underrepresented in the CAT version of the reading test. This represents a key limitation of CAT classification that should be highlighted for language testers.

The present study also has implications for curriculum and instruction. In educational contexts, assessing students' skills and processes in a particular field like language is the key to making informed decisions about the support and guidance the students need to reach their full potential. CAT as a dynamic type of assessment method has emerged as an alternative to traditional ones and is better adapted to the identification of learning difficulties and needs of the students. CAT can record a great amount of data about the examinees involving the sequence of actions taken by the student, the testing time, the successful and unsuccessful attempts made and the difficulty levels of each item. The possibility of obtaining all these data and to analyze them could offer valuable information which can be used for the intervention and improvement of the curriculum and instruction in a particular educational program. Rather than merely assessing the existing level of students' competencies, CAT also opens avenues for individualized instruction as it reveals what kind of support better can work for each student.

Implications for Further Research

The findings of this study point to various areas with potential for future research. The first area relates to the approach used in the study. Further studies are recommended to investigate the classification performance of CATs featuring multiple reading passages using the testlet approach. Another direction for research may be to implement the same methodological approach to investigate other

language skills to attain the optimum classification performance possible before administering a test or adopting a major change in the existing operational testing programs. Further research can also be done by implementing various classification methods and IRT models other than the binary model used in this study to evaluate testing programs based on the classification decisions they inform.

In the present study, only reading domain was taken into consideration. Listening domain was used as an external criteria. In addition to receptive skills like reading and listening, the applicability of CAT in writing and speaking domains as the productive language skills needs to be explored more. Developing CAT proficiency tests of productive skills is still not common as the computer softwares are not fully capable of evaluating written and spoken language.

The study was developed to examine the applicability of CAT in a language proficiency context and determine the impact of introducing a CAT on the classification accuracy and consistency of decisions made regarding English proficiency levels of students at a Turkish university. Overall, the study showed that regardless of the test termination rule, CAT improved levels of classification accuracy and consistency when a binary pass-fail decision was required. The role of language tests in Turkish higher education is increasingly prominent and test are often administered as a tool to make decisions that have important consequences for the examinees (Weiss, 2011). Such high stakes consequences make the accuracy and consistency of the classification estimations one of the critical issues to be considered before implementation regardless of test format. On a final note, as a result of the technological advancements and the ongoing challenges posed by the pandemic, how to continue assessment practices and ensure the growth and safety of the students has come under the spotlight to prevent this global health crisis from

becoming a generational educational crisis. The nature of how we test will need to align with the dramatic changes in how instruction has been delivered. It has become evident that we have to consider alternative models in various aspects of curriculum to tackle uncertainties, and we also have to be planning for the inevitability of change to how testing has been conducted over the years. In this respect, the present study could be enriched by exploring the preconceptions about testing. I believe that gaining literacy in new testing procedures like CAT is crucial and I hope that computerized adaptive testing will get the credit it deserves in language proficiency assessment.

Conclusion

The present study aimed to investigate the comparability of the CATs with reference to PBT and determine the impact of introducing a non-classification CAT on the classification accuracy and consistency of decisions relating to students' English language ability. To this end, a series of CAT simulations were conducted based on real examinee responses by using various test termination rules. Then, the accuracy and consistency of classification decisions were investigated with single and multiple cut-off points at different locations on ability continuum in different CAT simulations. Overall, results showed that the PBT and CAT ability estimates were not substantially different. Given that the test and item bank was designed for PBT with full fixed item administration, CAT yielded promising levels of classification accuracy and consistency when a binary pass-fail decision was required. This was the case regardless of the test termination rule. CAT also performed well with simultaneous use of multiple cut-off points. The results of this preliminary study demonstrated that non-classification CAT has high potential for classification purposes in pre-sessional English language programs. It is hoped that

the current study contributes to debates surrounding CAT in language testing and outlines a method of investigating the introduction of CAT in local contexts.

The findings obtained from the present study showed the potentialities of using regular CATs to classify students in an educational program. This preliminary study demonstrated that a regular CAT has the potential to give a similar level of classification performance to PBT with fewer items. We believe this is an important finding considering the limitations of administering PBT relating to time and the resources spent on such tests. With an initial investment into the item bank and the necessary facilities, using a regular CAT to classify learners can yield a similar classification performance to PBT. It also opens the door for more frequent assessment and early detection of misclassification, which will have a positive washback effect on teaching and learning practices in the English language preparatory program, which constitutes the context of the study.

Limitations

Although the present research presents clear conclusions pertaining to the applicability and classification performance of CAT for a language proficiency reading subtest, a number of limitations need to be acknowledged.

The first limitation is related to the size of the item bank. Having a real data set in the study can be considered an advantage, but the simulations were limited to 35 items. Item banks with larger number of items are inherently better at adapting CAT to each individual's ability level. Using a limited number of items yields less test information and consequently, decreases classification performance the size of the item bank used in the post-hoc simulations was small. Besides, it is important to bear in mind while interpreting the results of the study that the item bank was not designed for a typical CAT or CACT and therefore, does not cater for a large range

of ability. For this reason, the weaker levels of CA and CC at certain points on the ability continuum might also be attributed to this limitation. The study focused only on the B2 level of Common European Level of Reference for Languages, and this may limit the generalizability of the study. The test would be unable to make detailed distinctions between students that have substantially higher or lower levels of English.

Another limitation is related to the notion of CC as the level of agreement between parallel forms of the same test (Lee et al., 2008). Tests are considered parallel when their psychometric properties are identical (Boekkooi-Timminga, 1990). In a CAT context, a test has the potential to be different for each examinee because it is tailored to the ability level although tests can be statistically parallel. The items administered for each examinee may vary across two administrations based on his performance. Parallel adaptive tests have equal coverage of the content in the relevant domain. In order to obtain CC in CAT, “an item selection algorithm needs to be able to produce adaptive tests that are content balanced and not unduly overlapping” (Cheng & Morgan, 2012, p. 133). Two adaptive tests cannot be considered fully parallel if constraints like content coverage are controlled for. Although the present study considered statistical equivalency of tests, some non-statistical aspects like equivalency of content was partially controlled for, which may impede comparison made in ability estimation across various simulations. Administering CAT in a language comprehension test comprised of items targeting various reading processes may lead to an uneven representation of the item types among the test takers. This level of variation is important as it brings up the question whether stakeholders will accept that two statistically equivalent tests varying substantially in content measure the same reading construct.

The third limitation of the study is about the approach used in relation to the format of the test used as data. The reading subtest used for the study comprised of testlets, i.e., a bundle of items sharing the same reading comprehension passage as the common stimulus. For this study, testlet format in which items based on the same item stem or stimulus are administered collectively to create a single partial credit item per item stem might have been appropriate. This multistage testing approach was considered but not administered mainly because of the amount of obtainable data that would have reduced the number of reading passages and items available in the simulation and hence restricted the analysis. That's why, each item was considered independent rather than a member of a fixed bundle of items sharing a common reading passage. Statistical evidence indicated that the 35 items were independent, however, items that linked to the same reading passage often exhibit local dependence as any comprehension problem of the common stimulus is likely to impact the responses to the interrelated items (Eckes, 2017). The testlet approach presents a solution to this problem but only if the number of items available to create the CAT is sufficient.

REFERENCES

- ALTE Authoring Group (1998). *The multilingual glossary of language testing terms*. Cambridge University Press.
- Ames, A., & Penfield, R. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, 34(3), 39-48. <https://doi.org/10.1111/emip.12067>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment*, 8(1), 59-75. https://doi.org/10.1207/S15326977EA0801_04
- Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement. *Journal of Computerized Adaptive Testing*, 1(1), 1-18. <http://doi.org/10.7333/1212-0101001>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., & A. S. Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Marcel Dekker.

- Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-IBT writing tasks. *Language Testing*, 31(2), 241-259.
<https://doi.org/10.1177/0265532213509810>
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye- tracking. *Language Testing*, 30(4), 441-465.
<https://doi.org/10.1177/0265532212473244>
- Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Educational Testing Service.
- Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement*. 56(1), 121-146.
<https://doi.org/10.1111/jedm.12203>
- Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. In H. H. Goddard (Ed.), *Development of intelligence in children (the Binet-Simon Scale)* (pp. 274-329). Williams & Wilkins.
- Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
<https://doi.org/10.1007/BF02293801>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431-444. <https://doi.org/10.1177/014662168200600405>
- Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, 15, 129-145.
<https://doi.org/10.3102/10769986015002129>

- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Boussakuk, M., Bouchboua, A., El Ghazi, M., & Fattah, M. (2020). A fully individualized adaptive and intelligent educational hypermedia system: Details of CleverUniversity. *International Journal of Emerging Trends in Engineering Research*, 8(5), 1497-1502.
<https://doi.org/10.30534/ijeter/2020/04852020>
- Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts* [Final Report].
- Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229-255). Routledge.
- Breland, H., Lee, Y., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses (TOEFL Research Report No. RR-75)*. ETS.
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.
- Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling*, 58(4), 597-616.

- Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research*. 12(49), 61-80.
- Butler, F. A., Stevens, R., & Castellon, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 27-49). Yale University Press.
- Butterfield, M. S. (2016). *Comparing item selection methods in computerized adaptive testing using the rating scale model* [Unpublished doctoral dissertation]. The University of Texas at Austin.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*. 1(1), 1-47.
<https://doi.org/10.1093/applin/I.1.1>
- Cappelleri, J.C., Lundy, J.J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther*. 36(5), 648-62.
<https://doi.org/10.1016/j.clinthera.2014.04.006>
- Casillas, A., Robbins, S., McKinniss, T., Postlethwaite, B., & Oh, I. S. (2009). Using narrow facets of an integrity test to predict safety: A test validation study. *International Journal of Selection and Assessment*, 17(1), 119-125.
<http://doi.org/10.1111/j.1468-2389.2009.00456.x>
- Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer adaptive testing of reading proficiency*. Press Syndicate of the University of Cambridge.
- Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. British Council Monographs.

- Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3), 213-229.
<https://doi.org/10.1177/014662169602000303>
- Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222.
<https://doi.org/10.1177/01466219922031338>
- Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 40, 71-103. <https://www.jstor.org/stable/1435055>
- Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
<https://doi.org/10.1177/014662169602000303>
- Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
<https://doi.org/10.1177/01466219922031338>
- Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, 58, 569-595.
<https://doi.org/10.3102/1076998610381398>
- Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62(2), 369-383.
<https://doi.org/0.1348/000711008X304376>

- Cheng, Y., & Morgan, D. (2012). Classification accuracy and consistency of computerized adaptive testing. *Behavioral Research Methods*, 45(1), 132-142. <https://doi.org/10.3758/s13428-012-0237-6>
- Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language*, 8(2), 697-709
<https://nflrc.hawaii.edu/rfl/item/506>
- Christenson, S., Decker, D., Triezenberg, H., Yeseldyke, J., & Reschly, A. (2007). Consequences of high-stakes testing for students with and without disabilities. *Educational Policy*, 21(4), 662-690.
<https://doi.org/10.1177/0895904806289209>
- Choi, I. C. (1992). *An application of item response theory to language testing*. Peter Lang Publishing.
- Choi, I. C., & Bachman, L. F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing*, 9(1), 51-78. <https://doi.org/10.1177/026553229200900105>
- Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.
<https://doi.org/10.1191/0265532203lt258oa>
- Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 71, 37-53. <https://doi.org/10.1177/0013164410387338>
- Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, 33(6), 419-440.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

- College Board. (2003). *ACCUPLACER online technical manual*. The College Board.
- Council of Europe (2001). *The Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge University Press.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Cronbach, L. J. (1990). *Essentials of psychological testing*. Harper & Row.
<https://onlinelibrary.wiley.com/doi/epdf/10.1002/sce.3730350432>
- Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. *Handbook of Test Development*, 543-574. Routledge.
- Davey, T. (2011). *A guide to computer adaptive testing systems educational testing service for technical issues in large-scale assessment (TILSA)*. State Collaborative on Assessment and Student Standards (SCASS).
<https://files.eric.ed.gov/fulltext/ED543317.pdf>
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education*, 11(3), 231-253.
https://doi.org/10.1207/s15324818ame1103_2
- DeMars, C. E. (2010). *Item response theory: Understanding statistics measurement*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- DeMars, C. E. (2018). Item information function. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 899-903). SAGE. <https://dx.doi.org/10.4135/9781506326139.n360>

- Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton. CRC Press.
- Diao, H., & Sireci, S. G. (2018). Item response theory based methods for estimating classification accuracy and consistency. *Journal of Applied Testing Technology*, 19(1), 20-25. <https://doi.org/10.1111/j.1745-3984.2009.00096.x>.
- Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing*. Routledge. <https://doi.org/10.4324/9780429492242>
- Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology*, 2(2), 77-93. <http://dx.doi.org/10125/25044>
- Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, 34(3), 383-411. <https://doi.org/10.1177/0265532216672703>
- Educational Testing Service (ETS). (2009). *Guidelines for the assessment of English language learners*. ETS.
- Eignor, D. R., & Schaeffer, G. A. (1995). *Comparability studies for the GRE General CAT and the NCLEX using CAT* [Conference Presentation]. The Meeting of the National Council on Measurement in Education, San Francisco, United States.
- Eggen, T. J. H. (2009). Three-category adaptive classification testing. In W. van der Linden, & C. Glas (Eds.), *Elements of adaptive testing. Statistics for social and behavioral sciences* (pp. 373-387). Springer. https://doi.org/10.1007/978-0-387-85461-8_19
- Eggen, T. J. H. M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. In T. J. H. M. Eggen, & B. P.

Veldkamp (Eds.), *Psychometrics in Practice* (pp. 14-25). RCEC.

<http://dx.doi.org/10.3990/3.9789036533744>

Eggen, T. J. H. M. (2018). Multi-segment computerized adaptive testing for educational testing purposes. *Frontiers in Education*, 3(111), 1-8.

<https://doi.org/10.3389/feduc.2018.00111>

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60(5), 713-734.

<https://doi.org/10.1177/00131640021970862>

Eggen T. J. H. M., & Verschoor A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, 30(5), 379-393. <https://doi.org/10.1177/0146621606288890>

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational and Social Sciences*, 3(1), 57-75.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, 7(1), 19-29. <https://doi.org/10.1007/BF02288601>

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222, 309-368.

Fulcher, G., & Davidson, F. (Eds). (2012). *The Routledge handbook of language testing*. Routledge.

- Fulcher, G. (2012). Scoring performance tests. In G. Fulcher, & F. Davidson (Eds.) *The Routledge handbook of language testing* (pp. 378-392). Routledge.
- Garcia, G., McKoon, G., & August, D. (2006). Language and literacy assessment of language-minority students. In D. August, & T. Shanahan (Eds.), *Developing literacy in second-language learners* (pp. 597-624). Lawrence Erlbaum Associates.
- Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, 33, 157-192.
<https://doi.org/10.1111/j.1745-3984.1996.tb00487.x>
- Gitsaki, C., & Coombe, C. (Eds.). (2016). *Current issues in language evaluation, assessment and testing: Research and practice*. Cambridge Scholars Publishing.
- Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion and number of cut-scores. *Educational and Psychological Measurement*, 71(6), 1006-1022.
<https://doi.org/10.1177/0013164410393956>
- Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, 29, 433-456. <https://doi.org/10.1177/0146621605280072>
- Guan, Y. (2010). Methods for pattern classification. In Y. Zhang (Ed.), *New advances in machine learning* (pp. 49-74). IntechOpen.
<https://doi.org/10.5772/9378>

- Gushta, M. M. (2003). *Standard-setting issues in computerized-adaptive testing* [Paper Presentation]. The Annual Conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia.
- Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, 38(4), 558-579.
<https://doi.org/10.1177/0265532220979562>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.
- Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, 6(3), 535-556.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Han, K. T. (2009). *A gradual maximum information ratio approach to item selection in computerized adaptive testing* GMAC Research Report RR-09-07, June 25, 2009. USA.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345-359.
<http://www.jstor.org/stable/1434854>
- Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241-268). Cambridge University Press.
- Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, 11(2). 152-169.
<https://doi.org/10.1080/15434303.2014.902059>

- Harvey, R., & Hammer, A. (1999). Item response theory. *The Counseling Psychologist*, 27(3), 353-383. <https://doi.org/10.1177/0011000099273004>
- Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Instructional Topics in Educational Measurement*, 10(2), 33-41. <http://doi.org/10.1111/j.1745-3992.1991.tb00195.x>
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139-164. <http://dx.doi.org/10.1177/014662168500900204>
- Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20, 1-14. <https://doi.org/10.1177/014662169602000101>
- He, L., & Min, S. (2017). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly*, 14(2), 160-176. <https://doi.org/10.1080/15434303.2016.1162793>
- Helms, L. S. (1999, January 21-23). *Basic concepts in classical test theory: Tests aren't reliable, the nature of alpha, and reliability generalization as a meta-analytic method* [Conference presentation]. Annual Meeting, Southwest Educational Research Association, San Antonio, Texas, United States.
- Ho, T. H. (2010). *A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the generalized partial credit model* [Unpublished doctoral dissertation]. The University of Texas, Austin.
- Ho, T., & Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed-format adaptive test. *Applied Measurement in Education*, 25(4), 305-326. <https://doi.org/10.1080/08957347.2012.714686>

- Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8(3), 229-249.
<https://doi.org/10.1080/15434303.2011.565844>
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253-264.
<https://doi.org/10.1111/j.1745-3984.1976.tb00016.x>
- Hymes, D. H. (1971). On communicative competence. In J. Pride, & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Penguin.
- IACAT. 2016. *Research strategies in CAT*.
<http://iacat.org/content/research-strategies-cat>.
- Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, 4(45), 1-15.
<https://doi.org/10.3389/feduc.2019.00045>
- Johnson, M. (2018). Item information function. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 1-4). SAGE Publications. <https://dx.doi.org/10.4135/9781506326139.n694>
- Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished doctoral dissertation]. Middle East Technical University.
- Kalender, İ., & Berberoğlu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Educational Sciences: Theory & Practice*, 17, 573-596.
<http://doi.org/10.12738/estp.2017.2.0280>

- Kantarcıoğlu, E. (2012). *Relating an institutional proficiency examination to the CEFR: A case study* [Unpublished doctoral thesis]. University of Roehampton.
- Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması* [Yayınlanmamış doktora tezi]. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.
- Khoshsima, H., & Toroujeni, S. M. H. (2017). Computer-based testing: Score equivalence and testing administration mode preference in a comparative evaluation study. *International Journal of Emerging Technologies in Learning*, 12(10), 35-55. <https://doi.org/10.3991/ijet.v12i10.6875>
- Kim, S. Y., & Lee, W. C. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education*, 32(2), 97-115. <https://doi.org/10.1080/08957347.2019.1577246>
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6
- Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage Publications.
- Kormos, J. (2006). *Speech production and second language acquisition*. (Cognitive sciences and second language acquisition). Lawrence Erlbaum Associates.
- Kreiter, C. D., Ferguson, K., Gruppen, L. D. (1999). Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. *Academic Medicine*, 74(10), 1125-28.

<https://doi.org/10.1097/00001888-199910000-00016>

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research & Evaluation*, 20(18), 1-5. <https://doi.org/10.7275/43vm-p442>

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, 37(3), 226-241.

<https://doi.org/10.1177/0146621612471888>

Lawley, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology*, 33(3), 172-175.

<https://doi.org/10.1111/j.2044-8295.1943.tb01052.x>

Le, D. T. (2013). *Applying item response theory modeling in educational research* (Publication No. 13410) [Doctoral Dissertation, Iowa State University]. Doctor of Philosophy Graduate Theses and Dissertations.

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.) (2014). *Measuring L2 proficiency: Perspectives from SLA*. Multilingual Matters.

Lee, Y. J., Palazzo, D. J., Warnakulasooriya, R., & Pritchard, D. E. (2008). Measuring student learning with item response theory. *Physical Review Special Topics. Physics Education Research*, 4(1), 1-6.

<https://doi.org/10.1103/PhysRevSTPER.4.010102>

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47(1), 1-17. <https://doi.org/10.1111/j.1745-3984.2009.00096.x>

- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 367-386. <https://doi.org/10.1177/014662169001400404>
- Linacre, J. M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement*, 10(2), 157-169.
- Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, 41(7), 495-511. <https://doi.org/10.1177/0146621617707556>
- Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education*, 3(2), 115-141. https://doi.org/10.1207/s15324818ame0302_1
- Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11(31), 1-20. <https://doi.org/10.14507/epaa.v11n31.2003>
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching*. Pearson Education.
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, 73(2), 254-274. <https://doi.org/10.1177/0013164412453841>
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179-197. <https://doi.org/10.1111/j.1745-3984.1995.tb00462.x>

- Lord, M. F. (1970). Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139-183). Harper and Row.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157-162.
<https://www.jstor.org/stable/1434513>
- Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8(4), 453-461. <https://doi.org/10.1177/014662168400800409>
- Lu P., Zhou D., Qin S., Cong X., Zhong S. (2012). The study of item selection method in CAT. In Li Z., Li X., Liu Y., Cai Z. (Eds.), *Computational Intelligence and Intelligent Systems. ISICA 2012. Communications in Computer and Information Science* (pp. 413-415). Springer.
https://doi.org/10.1007/978-3-642-34289-9_45
- Matteucci, M., & Veldkamp, B. P. (2013). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Applications*, 22(2), 243-267.
<https://doi.org/10.1007/s10260-012-0216-1>
- Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M., von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background,*

technical issues, and methods of data analysis (pp.230-260). Chapman and Hall/CRC.

McDonald, R. P. (1999). *Test theory: A unified treatment*. LEA.

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, 36(1), 101-123.

<https://doi.org/10.1177/0265532217725776>.

Morris, S. B., Bass, M., Howard, E., & Neapolitan, R. E. (2020). Stopping rules for computer adaptive testing when item banks have nonuniform information. *International Journal of Testing*, 20(2), 46-168.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 47(4), 355-366.

<https://doi.org/10.1037/h0059934>

Mosier, C. I. (1941). Psychophysics and mental test theory (II. The constant process). *Psychological Review*, 48(3), 235-249. <https://doi.org/10.1037/h0055909>

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data* [computer program]. Scientific Software, Inc.

Nunnally, J.C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory*, 3, 248-292.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. Longman.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, 30(3), 363-380.

<https://doi.org/10.1177/0265532213480336>

Ostini, R., Finkelman, M., & Nering, M. (2015). Selecting among polytomous IRT models. In S. P. Reise, & D. A. Revicki (Eds.), *Multivariate applications*

- series. *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 285-304). Routledge/Taylor & Francis Group.
- O'Sullivan, B. (2016). Validity: What is it and who is it for? In Y. N. Leung (Ed.), *Epoch making in English teaching and learning: Evolution, innovation, and revolution* (pp. 157-175). Crane Publishing Company Ltd.
- Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education*, 19(1), 1-20. https://doi.org/10.1207/s15324818ame1901_1
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.
- Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.
- Reise, R. S., & D. A. Revicki (Eds.). (2013). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge. <https://doi.org/10.4324/9781315736013>
- Reynolds, T., Perkins, K., & Brutten, S. (1994). A comparative item analysis study of a language testing instrument. *Language Testing*, 11, 1-14. <https://doi.org/10.1177/026553229401100102>
- Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1(2), 33-49.
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(14). <https://doi.org/10.7275/an9m-2035>
- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13). <https://doi.org/10.7275/56a5-6b14>

- Rudner, L. M., & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *GMAC Research Reports*, 11-01.
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63-84. <https://doi.org/10.1177/0013164404273942>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(4), 1-97.
- Samsudin, M. A., Chut, T. S., Ismail, M. E., & Ahmad, N. J. (2020). A calibrated item bank for computerized adaptive testing in measuring science TIMSS performance. *Eurasia Journal of Mathematics Science and Technology Education*, 16(7), 1-15. <https://doi.org/10.29333/ejmste/8259>
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computer adaptive testing: From inquiry to operation*. APA.
- Schuster, C. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, 36(6), 720-735. <https://doi.org/10.3102/1076998610396890>
- Segalowitz, N. (2010). *The cognitive bases of second language fluency*. Routledge.
- Segall, D. O. (1995, April). *Equating the CAT-ASVAB: Experiences and lessons learned* [Paper presentation]. The Meeting of the National Council on Measurement in Education, San Francisco.
- Selvi, A. (2014). The medium-of-instruction debate in Turkey: Oscillating between national ideas and bilingual ideals. *Current Issues in Language Planning*, 15(2), 133-152. <https://doi.org/10.1080/14664208.2014.898357>
- Sheerin S. (1987). Listening comprehension: Teaching or testing? *ELT Journal*, 41(2), 126-131. <https://doi.org/10.1093/elt/41.2.126>

- Sijtsma, K. (2015). Playing with data—Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*.
<http://doi.org/10.1007/s11336-015-9446-0>
- Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (ETS Research Report No. RR-11-29). ETS.
- Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Report No. ACT-RR-93-7). American College Testing.
- Spray J. A., & Reckase M. D. (1994, April 5-7). *The selection of test items for decision making with a computer adaptive test* [Paper presentation]. National Meeting of the National Council on Measurement in Education, New Orleans, LA. <https://eric.ed.gov/?id=ED372078>
- Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs*. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review*, 36, 263-277.
<https://doi.org/10.1111/j.1464-0597.1987.tb01191.x>
- Stone, C.A., Weissman, A., & Lane, S. (2005). The consistency of student proficiency classifications under competing IRT models for a state assessment program. *Educational Assessment*, 10(2), 125-146.
https://doi.org/10.1207/s15326977ea1002_3
- Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Kuram ve*

Uygulamada Egitim Bilimleri, 15(6), 1585-1595.

<https://doi.org/10.12738/estp.2015.6.0102>

Tejada, A. J. R., & Rojas, O. M. L. (2005). Application of an IRT polychromous model for measuring health related quality of life. *Social Indicators Research*, 74, 369-394. <http://doi.org/10.1007/s11205-004-3232-1>

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101-133). Lawrence Erlbaum.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, 12(1).

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, 69(5), 778-793.

<https://doi.org/10.1177/0013164408324460>

Thompson, N. A., & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC conference on computerized adaptive testing*.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, 16(1). <https://doi.org/10.7275/wqzt-9427>

Triola, M. F. (2007). *Elementary Statistics* (10th Ed.). Addison-Wesley.

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, 97, 69-85.

<https://doi.org/10.1016/j.compedu.2016.02.018>

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. Springer.

- van der Linden W. J., & Pashley P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden, & C. A. W. Glas CAW (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-288). Kluwer Academic Publishers.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer-Verlag.
- van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. Springer.
- van Steendam, E. , Tillema, M., Rijlaarsdam, G., & Bergh, H. (Eds.). (2012). *Measuring writing: Recent insights into theory, methodology and practice*. Emerald Group Publishing Limited.
- van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing *Applied Psychological Measurement*, 26, 393-411.
<https://doi.org/10.1177/014662102237796>
- Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 207-214). Springer.
- Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22(2), 203-226. <https://doi.org/10.3102/10769986022002203>
- Verhoeven, L., & de Jong, J. H. A. L. (1991). *Modelling and assessing language proficiency: The construct of language proficiency*. John Benjamins Publishing.

- Verschoor, A. J., & Straetmans, G. J. J. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 137-149). Statistics for Social and Behavioral Sciences. Springer
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Thissen, D., Steinberg, L., & Thissen, D. (Eds.). (2000). *Computerized adaptive testing: A primer*. Routledge.
- Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer et al. (Ed.), *Computer adaptive testing: A primer* (pp. 271-99). Lawrence Erlbaum Associates.
- Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, 12(4), 339-368.
<https://doi.org/10.3102/10769986012004339>
- Walczak, A. (2015). Computer-adaptive-testing. *Research Notes* 59, 35-39.
- Wang, T., & Kolen, M.J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38(1), 19-49. <https://doi.org/10.1111/j.1745-3984.2001.tb01115.x>
- Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurement and Research Service Bulletin*, 13, 1-7.
- Wang, T., & Vispoel, W. P. (2005). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(2), 109-135. <https://doi.org/10.1111/j.1745-3984.1998.tb00530.x>
- Wang, X., Fluegge, L., & Luecht, R. (2012). *A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations* [Paper

presentation]. Annual Meeting of the National Council on Measurement in Education. Vancouver, British Columbia, Canada

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, 23(3), 263-278.

<https://doi.org/10.1177/01466219922031383>

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, 54, 427-450.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas assessment of knowledge and skills* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education 2006, San Francisco, CA.

https://images.pearsonassessments.com/images/tmrs/Score_Comparability_of_Online_and_Paper_Administrations_of_TAKS_03_26_06_final.pdf

Weir, C., Hawkey, R., Green, T., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. British Council/IDP Australia IELTS Research Reports, 9(4), 157-189.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473-492.

<https://doi.org/10.1177/014662168200600408>

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361-375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>

Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in*

Counseling and Development, 37(2), 70-84.

<https://doi.org/10.1080/07481756.2004.11909751>

Weiss, D. J. (2005). Manual for POSTSIM: Post hoc simulation of computerized adaptive testing (Version 2.0) [Computer software]. Assessment Systems Corporation.

Weiss, D. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1-27.

Weissman, A. (2004, April). *Mutual information item selection in multiple-category classification CAT* [Paper presentation]. The Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

<https://doi.org/10.1177/0013164406288164>

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, 67(1), 41-58.

<https://doi.org/10.1177/0013164406288164>

Wiersma, W., & Jurs, S. G. (Eds). (1990). *Educational measurement and testing*. Allyn and Bacon.

Wolf, M. K., Kao, J. Griffin, N., Herman, J. L., Bachman, P. L., Chang, S., & Farnsworth, T. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses - Practice review* (CSE Tech. Rep. No. 732). University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores* (TOEFL Research Report No. RR-72). ETS.

- Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement*, 66(4), 545-564. <https://doi.org/10.1177/0013164405284031>
- Yao, L. (2019). Item selection methods for computer adaptive testing with passages. *Frontiers in Psychology*. 10(240), 1-10. <https://doi.org/10.3389/fpsyg.2019.00240>
- Yeğin, P. O. (2003). *The predictive validity of Başkent university proficiency exam (BUEPE) through the use of the three-parameter IRT model's ability estimates* [Unpublished master's thesis]. Middle East Technical University. <https://etd.lib.metu.edu.tr/upload/1013507/index>
- Yen, W., & Fitzpatrick, A. R. (2008). Item response theory. In R. Brennan (Ed.), *Educational measurement* (pp. 111-153). Praeger.
- Yi, Q., Wang, T., & Ban, J. C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, 38, 267-292.
- Young, F., Shermis, M. D., Brutton, S., & Perkins, K. (1996). From conventional to computer adaptive testing of ESL reading comprehension. *System*, 24(1), 32-40. [https://doi.org/10.1016/0346-251X\(95\)00051-K](https://doi.org/10.1016/0346-251X(95)00051-K).
- Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students*. Development Associates, Special Issues Analysis Center.
- Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27(1), 119-140. <https://doi.org/10.1177/0265532209347363>

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Scientific Software, Inc.

Zopluoglu, C., & Davenport, E. C. (2017). A note on using eigenvalues in dimensionality assessment, *Practical Assessment, Research, and Evaluation*. 22(7), 1-10. <https://doi.org/10.7275/zh1k-zk32>



Appendix A

Correctly Classified Examinees into the Same Decile by SE-based CATs

Table A

Percentages of Correctly Classified Examinees into the Same Decile by SE-based CATs (One Ability Group Above or Below by PBT)

Deciles	SE threshold test termination rule				
	0.10	0.20	0.30	0.40	0.50
10%	100 (0%)	100 (0%)	100 (0%)	82 (18%)	73 (24%)
20%	100 (0%)	99 (0%)	99 (1%)	62 (36%)	39 (47%)
30%	92 (2%)	92 (1%)	83 (14%)	46 (45%)	31 (48%)
40%	100 (0%)	100 (0%)	82 (26%)	36 (49%)	19 (49%)
50%	91 (1%)	91 (1%)	83 (12%)	49 (49%)	33 (45%)
60%	99 (1%)	99 (1%)	99 (1%)	62 (30%)	36 (38%)
70%	99 (1%)	99 (1%)	99 (1%)	68 (31%)	38 (43%)
80%	100 (0%)	99 (0%)	97 (1%)	79 (18%)	39 (50%)
90%	100 (0%)	100 (0%)	60 (40%)	97 (0%)	56 (37%)
100%	97 (1%)	99 (1%)	100 (0%)	97 (1%)	68 (13%)

Appendix B

Correctly Classified Examinees into the Same Decile by FL CATs

Table B

Percentage of Correctly Classified Examinees into the Same Decile by FL CATs (% of Examinees Classified into One Ability Group Above or Below)

Deciles	Fixed-length test termination rule				
	10 items	15 items	20 items	25 items	30 items
10%	79 (20%)	80 (20%)	87 (13%)	96 (4%)	96 (4%)
20%	44 (40%)	51 (48%)	66 (33%)	79 (20%)	87 (13%)
30%	32 (50%)	41 (49%)	56 (43%)	60 (39%)	80 (20%)
40%	21 (48%)	35 (48%)	55 (40%)	58 (36%)	76 (24%)
50%	38 (42%)	39 (46%)	47 (43%)	49 (43%)	72 (25%)
60%	28 (43%)	33 (49%)	39 (49%)	52 (45%)	66 (32%)
70%	33 (46%)	41 (46%)	43 (48%)	58 (40%)	74 (23%)
80%	40 (44%)	45 (46%)	52 (44%)	65 (33%)	78 (20%)
90%	42 (48%)	50 (48%)	63 (35%)	70 (27%)	83 (16%)
100%	67 (30%)	75 (23%)	84 (15%)	87 (10%)	92 (7%)

Appendix C

Details of the Twenty-Five Most Frequently Used Items in CAT Simulations

Table C

Categories of the Twenty-Five Most Frequently Used Items in CAT Simulations

Test Termination Rule																	
SE03			SE04			SE05			FL30			FL25			FL20		
Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count	Item #	Domain	Count
33	A	1182	33	A	1182	29	B	1155	33	A	1182	33	A	1182	33	A	1182
35	A	1182	35	A	1182	14	C	1097	35	A	1182	35	A	1182	29	B	1182
12	A	1182	28	A	1182	28	A	1079	12	A	1182	12	A	1182	4	B	1161
28	A	1182	11	B	1182	11	B	923	18	A	1182	4	B	1182	35	A	1155
18	A	1182	14	C	1182	27	B	795	20	A	1182	23	B	1182	14	C	1126
16	A	1182	27	B	1178	33	A	759	34	A	1182	15	B	1180	23	B	1078
19	A	1182	29	B	1177	18	A	736	4	B	1182	21	B	1178	12	A	1045
17	A	1182	12	A	1115	19	A	718	23	B	1182	29	B	1174	21	B	1014
20	A	1182	17	A	1108	16	A	685	21	B	1182	9	C	1171	28	A	977

29	B	1182	23	B	1084	23	B	679	11	B	1182	14	C	1160	18	A	921
4	B	1182	20	A	1073	21	B	606	15	B	1182	20	A	1126	9	C	921
23	B	1182	3	C	1064	35	A	588	14	C	1182	6	D	1122	11	B	913
21	B	1182	19	A	1060	30	B	570	9	C	1182	18	A	1081	15	B	898
11	B	1182	4	B	991	9	C	523	7	C	1182	7	C	1070	16	A	826
15	B	1182	18	A	966	4	B	444	6	D	1182	28	A	1052	27	B	792
27	B	1182	15	B	964	32	B	442	29	B	1179	11	B	1002	19	A	741
30	B	1182	21	B	955	5	A	422	28	A	1168	16	A	960	17	A	683
13	B	1182	13	B	887	3	C	416	24	B	1156	27	B	889	30	B	679
14	C	1182	16	A	871	1	A	403	17	A	1150	30	B	843	20	A	653
9	C	1182	9	C	836	13	B	398	3	C	1109	3	C	778	3	C	632
3	C	1182	7	C	817	17	A	387	16	A	1105	19	A	768	32	B	559
7	C	1182	30	B	783	25	D	385	10	B	1099	17	A	742	13	B	546
6	D	1182	2	B	762	2	B	363	32	B	1079	34	A	682	25	D	520
31	E	1182	6	D	743	22	A	350	30	B	1004	32	B	660	6	D	496
34	A	1179	34	A	729	24	B	305	27	B	988	13	B	614	22	A	488

Note. A= propositional inference; B= paraphrase of supporting detail; C= pragmatic inference; D= meaning of unknown lexis in context; E= textual cohesion. Count refers to the number of times that the item features in a test

Appendix D

Details of the Items in PBT

Table D

Item Parameters and Categories in PBT

Item No.	Item Category	Item Difficulty	Item Discrimination
1	A	0,46	0,26
2	B	0,84	0,70
3	C	0,72	0,65
4	B	0,55	0,47
5	A	0,26	0,36
6	D	0,44	0,42
7	C	0,59	0,43
8	A	0,56	0,28
9	C	0,43	0,51
10	B	0,67	0,39
11	B	0,66	0,78
12	A	0,60	0,49
13	B	0,80	0,87
14	C	0,56	0,84
15	B	0,60	0,47
16	A	0,35	0,62
17	A	0,71	0,68
18	A	0,40	0,60
19	A	0,74	1,07
20	A	0,66	0,54
21	B	0,44	0,53
22	A	0,31	0,48
23	B	0,46	0,54
24	B	0,37	0,43
25	D	0,34	0,47

Note. A= propositional inference; B=paraphrase of supporting detail; C=pragmatic inference; D= meaning of unknown lexis in context; E= textual cohesion

VITA

Elif Kaya

CONTACT INFORMATION

English Language Preparatory School
Bilkent University
East Campus/ N Building

EDUCATION

- 2021, Ph.D. Curriculum and Instruction, Bilkent University, Ankara, Turkey. (under the supervision of Asst. Prof. Dr. İlker Kalender). The title of my PhD dissertation: *A Comparability and Classification Analysis of Computerized Adaptive and Conventional Paper-Based Versions of an English Language Proficiency Reading Subtest* (Current GPA : 3.71)
- 2010, M.A. Management in Education, Bilkent University, Ankara, Turkey (GPA: 3.98). MA Project Title: *Exploring Students' Perceptions of Outside Class Studies*
- 2009, Level 7 Diploma in Teaching English to Speakers of Other Languages (DELTA), University of Cambridge. *Courses taken at Bilkent University School of English Language, Ankara, Turkey*
- 2005, In-service Certificate in English Language Teaching (ICELT), University of Cambridge. *Courses taken at Bilkent University School of English Language, Ankara, Turkey*
- 2003, B.A. Foreign Language Education, Middle East Technical University, Ankara, Turkey (GPA: 3.57).
- 2003, Minor: Political Science, Middle East Technical University, Ankara, Turkey.

EMPLOYMENT

- August, 2004-Present, Instructor of English, Bilkent University English Language Preparatory Programme, Bilkent University, Ankara.

CONFERENCES/ SYMPOSIUMS/ WORKSHOPS ATTENDED

- December, 2017, Bilkent University, Graduate School of Education, Graduate Seminar 2 (Presenter). *A Comparability Analysis of Scores from Computerized Adaptive and Paper-Pencil Format of a Language Proficiency Test*. Bilkent University, Ankara, Turkey.
- October, 2015, International Higher Education (IHEC) Conference (Presenter). *An Effective Classification Approach for Language Proficiency Tests based on Computer Adaptive Testing*. Boğaziçi University, İstanbul, Turkey.
- April, 2013, IATEFL Conference (Presenter). *Learners' Perceptions of Outside Class Studies*. Liverpool, England.
- June, 2012, Bilkent University CIDER Research Symposium, Ankara.
- June, 2012, Bilkent University BUSEL EAP Conference, Ankara

- June, 2011, BUSEL Research Symposium (Presenter): Exploring the students' perceptions of their outside-class learning experiences, Bilkent University, Ankara.
- June, 2010, EAP in University Settings: Teacher and Learner Competencies, Bilkent University, Ankara.
- June, 2009, Critical Thinking Workshop for Preparatory Program Instructors, Bilkent University, Ankara.
- June, 2007, 10th International ELT Conference. The Future of ELT: A Quest for Common Standards & Approaches, Bilkent University, Ankara.

PUBLICATIONS

- Kaya, E., O'Grady, S., & Kalender, İ. (In-press). IRT-based classification analysis of an English language reading proficiency subtest. *Language Testing*. (Research Article)
- Kaya, E., & Kalender, İ. (2018). Use of Computerized Adaptive Testing in Class-level Language Assessment. In Aksit, T., Mengü, H. I., & Turner, R. (Eds.), *Bridging teaching, learning and assessment in the English language classrooms*. (pp. 73-83). Cambridge Scholars Publishing. (Book Chapter).
- Kaya, E. & Kalender, İ. (2016). Concurrent Use of Computerized Adaptive Testing in Class-Level Language Assessment (Paper presentation) 4th International Bilkent University School of English Language Conference, Ankara, Turkey. (Conference Paper)
- Kaya, E. & Kalender, İ. (2015). An Efficient Classification Approach for Language Proficiency Tests based on Computerized Adaptive Testing. (Paper presentation). The 1st International Higher Education Studies Conference, Ankara, Turkey. (Conference Paper)

PROFESSIONAL INTERESTS

- Curriculum Development
- Educational Assessment and Measurement
- Computerized Testing
- Language Teaching