DEVELOPING A SCALE ON EXAMINEES' ATTITUDES TOWARD COMPUTERIZED ADAPTIVE TESTING: APPLYING TECHNOLOGY ACCEPTANCE MODEL

by

Beyza Arpacıoğlu

B.S., Physics Education, Boğaziçi University, 2021

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Secondary School Science and Mathematics Education

Boğaziçi University

2024

ACKNOWLEDGEMENTS

First of all, I would like to express my heartfelt appreciation to my esteemed supervisor Assoc. Prof. Serkan Arıkan for unflagging support, unforgettable patience and beliefs in me and current study. He always standed behind me in all stages, it was incredibly valuable for me, I feel very fortunate to have had the chance to work with him.

I am also grateful to my thesis committee members, Prof. Emine Erktin and Assoc. Prof. Eren Can Aybek for generously dedicating their time and providing constructive feedback to the study.

I want to express my eternal gratitude my parents, Turafiye and Bülent Arpacıoğlu, and my sister Berra Arpacıoğlu, for their unconditional support, patience, infinite love and trust throughout current study and my whole life. I am also gratitude my dear friends, Büşra Ertürk, Mehtap Kızılkaya and Mukaddes Betül Reyhan who have enveloped me with their love and unwavering support.

Lastly, I'm also deeply appreciative of İlke Öztürk, Global Learning and Experience Leader of Borusan Cat, for her understanding and unflagging support during my academic journey. This journey was peerless experience, filled with perseverance and passion, for me. My sincere thanks to everyone who has been a part of it.

ABSTRACT

DEVELOPING A SCALE ON EXAMINEES' ATTITUDES TOWARD COMPUTERIZED ADAPTIVE TESTING: APPLYING TECHNOLOGY ACCEPTANCE MODEL

Implementations in measurement are greatly affected by and quickly adapt to recent advances in the field of information and communication technology (ICT) (Bennett, 2002; Linn, 1993). Particularly, the noticeable growth in computer usage by the end of the 1980s has demonstrated its impact on computer-based administrations in measurement and evaluation (Van der Linden *et al.*, 2000; McDonald, 2002). Correspondingly, the usage of Computerized Adaptive Testing (CAT) has increased significantly and more examinees experience CAT. Thus, understanding examinees' attitudes toward CAT is essential for further CAT implementations to be more successful and to be used widely in educational systems (Lilley *et al.*, 2011; Özbaşı, 2016). However, there are a limited number of studies that investigating examinees' attitudes toward CAT in the literature and there are a few instruments that measures examinees' attitudes toward CAT in educational testing. Thus, the purpose of the current study is to develop a scale to measure examinees' attitudes toward computerized adaptive testing. Technology Acceptance Model (TAM) provides a theoretical framework for developing dimensions of this scale. Proposed model in current study has three dimensions: perceived ease of use, perceived usefulness and perceived enjoyment. The initial version of the scale data exhibited a good fit with measurement model (CFI=0.970, TLI=0.964, RMSEA=0.067). Following revisions, the final version of the scale data showed a good degree of compatibility with proposed measurement model (CFI= 0.969, TLI= 0.959, RMSEA=0.067).

ÖZET

BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TESTLERE KARŞI ÖĞRENCİ TUTUMLARI: TEKNOLOJİ KABUL MODELİ İLE ÖLÇEK GELİŞTİRME ÇALIŞMASI

Ölçmedeki uygulamalar, bilgi ve iletişim teknolojisi alanındaki son gelişmelerden büyük ölçüde etkilenmekte ve bunlara hızla uyum sağlamaktadır (Bennett, 2002; Linn, 1993). Özellikle 1980'li yılların sonuna doğru bilgisayar kullanımındaki gözle görülür artış, ölçme ve değerlendirmede, bilgisayar tabanlı testlerin kullanımına da etkisini göstermiştir (Van der Linden *et al.*, 2000; McDonald, 2002). Buna paralel olarak, Bilgisayar Ortamında Bireye Uyarlanmış Testlerin (BOBUT) kullanımı önemli ölçüde artmış ve daha fazla kişi BOBUT deneyimini yaşamaya başlamıştır. Bu nedenle, gelecekteki BOBUT uygulamalarının daha başarılı olabilmesi ve eğitim sistemlerinde yaygın olarak kullanılabilmesi için sınava girenlerin BOBUT'a yönelik tutumlarının anlaşılması önem kazanmıştır (Lilley *et al.*, 2011; Özbaşı, 2016). Ancak literatürde BOBUT deneyimini yaşayanların BOBUT'a yönelik tutumlarını araştıran sınırlı sayıda çalışma bulunmakla beraber, BOBUT'a yönelik tutumlarını ölçmek için kullanılabilecek yalnızca birkaç araç bulunmaktadır. Dolayısıyla bu çalışmanın amacı, BOBUT katılımcılarının BOBUT'a yönelik tutumlarını ölçecek bir ölçek geliştirmektir. Teknoloji Kabul Modeli (TAM), bu ölçeğin boyutlarının geliştirilmesine yönelik kuramsal çerçeveyi sağlamaktadır. Bu çalışmada önerilen modelin üç boyutu vardır: algılanan kullanım kolaylığı, algılanan fayda ve algılanan keyif alma. Pilot uygulamadan elde edilen veriler ölçüm modeliyle iyi bir uyum sergilemiştir (CFI= 0,970, TLI= 0,964, RMSEA= 0,067). Ölçeğe son halinin verilmesinin ardından yapılan asıl uygulama verileri de benzer şekilde ölçüm modeliyle iyi düzeyde uyumluluk göstermiştir (CFI= 0,969, TLI= 0,959, RMSEA= 0,067).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AERA	American Educational Research Association
APA	American Psychological Association
ASVAB	Armed Services Vocational Aptitude Battery
ATT	Attitude Toward Using
BI	Behavioral Intention
CAT	Computerized Adaptive Testing
CBA	Computer Based Assessment
CBAAM	Computer Based Assessment Acceptance Model
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
EAP	Expected A Posteriori
GPCM	Generalized Partial Credit Model
GPM	Graded Response Model
GRE	Graduate Records Examination
ICC	Item Characteristic Curve
ICT	Information and Communication Technology
IRT	Item Response Theory
MAP	Maximum A Posteriori
MFI	Maximum Fisher Information
NCME	National Council on Measurement in Education
OECD	Organization for Economic Cooperation and Development
P and P	Paper and Pencil
PE	Perceived Enjoyment
PEOU	Perceived Ease of Use
PISA	Programme for International Student Assessment
PU	Perceived Usefulness
TAM	Technology Acceptance Model
TOEFL	Test of English as a Foreign Language

TPB	Theory of Planned Behavior
UTAUT	The Unified Theory of Acceptance and Usage of Technology

1. INTRODUCTION

The first chapter introduces background, purpose, significance and research questions of the study respectively.

1.1. Background of the Study

Measurement is an essential activity in all branches of science, involving the behavioral and social sciences (DeVellis, 2017). In this sense, measurement is also a fundamental component of education. Educational testing is among the most significant contributors of behavioral science to society with numerous critical functions (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), 1999). Especially, four distinct functions are noteworthy. Firstly, measurement plays a vital role in wide variety of decisions about students such as admission to college and high school graduation there or more references should be cited as (Thorndike *et al.*, 1991) so that your multiple references should be (Linn, 1993; Thorndike *et al.*, 1991). The second role is gaining insight about achievement of different student groups broadly. Rather than following each student academically at close range, good estimations can be obtained via administering subsets of items to samples of students as in TIMSS, PISA and PIRLS. Test results of students are used to make judgments about the adequacy of education system. Thirdly, accountability function which highlights that educational tests are used as critical educational reform tool. Final major function is formative decision making by classroom teachers which refers to providing continual feedback that can be used by teachers and students to enhance instruction (Linn, 1993). In summary, it is indisputable that data obtained through measurement reflects the strengths and weaknesses of education system as well as knowledge about students and teachers, then offers the opportunity to enhance the quality and efficiency of education.

In recent years, there have been many advances in the field of information and communication technology (ICT). Implementations and preferences in measurement are also most influenced by and adapt to these alterations rapidly (Bennett, 2002; Linn, 1993). In particular, the significant increase in the computer use at the end of the 1980s has demonstrated its influence on computer-based administrations in the field of measurement and evaluation (Van der Linden *et al.*, 2000; McDonald, 2002). There are two ways a test might be implemented via computers: Computer - Based Assessment (CBA) and Computerized Adaptive Testing (CAT). The former can be defined as the computerized version of traditional pencil and paper tests (P and P). Examinees are given questions in the previously defined order. The main difference of CBA from P and P is that examinees take tests on a computer rather than paper (Burke *et al.*, 1987; Davey, 2011; Hoffman *et al.*, 1976). The latter is, Computerized Adaptive Testing, and explained more comprehensively below.

1.1.1. Computerized Adaptive Testing

Computerized adaptive testing refers to tailored, response-contingent, programmed, individualized, branched and sequential testing. CAT uses statistical models such as item response theory (IRT) by selecting the next item depending on the response to the most recent one. When an examinee answers the question incorrectly, next item delivered will be easier; when the response is correct, next item will be more difficult (Weiss, 1985). In order to better understand the working principle of CAT, application process is summarized (Davey, 2011; Hambleton *et al.*, 1984; Lord *et al.*, 1988) and illustrated in Figure 1.1 below (Davey, 2011):

- Starting Rule. There are various methods regarding the selection of the first item. Starting test with medium-difficulty item is one of the most common approaches if there is no prior information about the examinee's level.
- Item Selection. Before CAT implementation, items should be piloted and item parameters should be calculated. Item pool consisting of a large number of questions with various difficulty values should be generated so that CAT algorithm

can work accordingly.

The most commonly used method in item selection is the Maximum Fisher Information (MFI). In this method, the next item expected to provide the highest information for the estimated ability level is selected from the item pool.

- Ability Estimation. Following each question, CAT updates ability estimation of examinees using different mathematical methods such as Expected a Posteriori (EAP) and Maximum a Posteriori (MAP).
- Stopping Rule. Until termination criterion can be met, previous steps are repeated. Stopping rule might be falling below a certain user-specified value of standard error of measurement or item number.

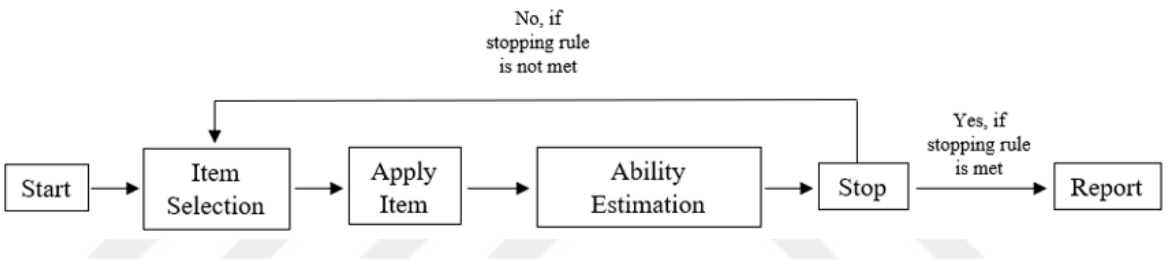

Figure 1.1. A CAT Process (Davey, 2011).

In CAT, examinees' ability levels can be determined with fewer questions which leads to shorter testing time. As examinees engaged with items according to their level, they do not spend time answering very easy or very difficult questions for their levels (Embretson *et al.*, 2000; Mcglohen *et al.*, 2008; Weiss *et al.*, 1984). Besides its advantages, CAT might have disadvantages such as the effects of the level of technological literacy on test scores and high cost (Çıkrıkçı *et al.*, 1999; Embretson *et al.*, 2000; Hambleton *et al.*, 1991; Weiss, 1985).

The use of CAT has grown up globally with a notable increase in computer usage. CAT has been existing for 50 years in the literature but studies have been more heavily concentrated on it in the last 20 years (Blais *et al.*, 2002; Wainer, *et al.*, 1990; Weiss, 2004). Despite the fact that there have been several studies on a variety of CAT-related topics in the literature, there is a limited number of studies related to examinees' attitudes towards CAT. There is scarcely any assessment tool that measures

examinees' attitudes toward CAT. However, to be informed of attitudes of test-takers toward CAT is an irrefutable requirement. CAT implementations might progress solely by feeding on attitudes of the test-takers and become more preferable over P and P test or CBA. Thus, teachers, students, academics and policy makers might recognize the need to incorporate CAT in their assessment procedures (Daphine *et al.*, 2020). At the end of this chain, understanding attitudes of examinees toward CAT might lay the foundations of educational reform.

1.1.2. Technology Acceptance Model

The Technology Acceptance Model (TAM), a theory of information systems presented by (Davis, 1989), models how users come to accept or reject and use a technology in different contexts such as e-learning, web-based assessments, and computer-based assessments (Davis *et al.*, 1989). In other words, TAM is one of the existing models to investigate end-users' behavioral intention towards technology use and frequently used as a framework to develop scales (Davis *et al.*, 1989). Causal relationships between system design features, perceived usefulness, perceived ease of use, behavioral intention, attitude toward using and actual usage behavior are specified by TAM (Davis, 1989). According to the model, when individuals are exposed to new technology, a variety of factors affect their choice of how and when to use it. System design features, often called as external factors, refer to characteristics of the information system used. The model suggests two main individual beliefs as perceived ease of use and perceived usefulness impacting an individual's attitude toward using new technology. The behavioral intention is influenced by the attitude which refers to the overall judgment of users about new technology. The actual system usage behavior is the ultimate point where people use the technology. TAM, a well-accepted model, has been frequently used in and expanded to different studies and has an important place in the literature (Alkış, 2010; Dizon, 2016; Lee*et al.*, 2009; Ratna *et al.*, 2015; Teo *et al.*, 2008; Terzis *et al.*, 2011; Terzis *et al.*, 2011).

1.2. Significance of the Study

Applications of CAT have been getting more widespread, even in high-stakes admission tests such as Armed Services Vocational Aptitude Battery (ASVAB) (McBride, 2001), Test of English as a Foreign Language (TOEFL) (Glas *et al.*, 2003), Graduate Records Examination (GRE) (Wainer *et al.*, 1990) and Programme for International Student Assessment (PISA) (Schleicher, 2019). Correspondingly, the number of examinees who experience CAT has been increasing all around the world. Understanding these examinees' attitudes toward CAT is an essential need for future CAT implementations to be more effective (Lilley *et al.*, 2011; Özbaşı, 2016). There are, however, a limited number of studies in the literature that investigate examinees' attitudes toward CAT (e.g., Daphine *et al.*, 2020; English *et al.*, 1977; Legg *et al.*, 1992; Moe *et al.*, 1988; Özbaşı, 2016; Schmidt *et al.*, 1978; Tonidandel *et al.*, 2000; Vispoel *et al.*, 1994). Thus, the current study aimed to develop a scale to measure examinees' attitudes toward computerized adaptive testing. The findings of current study are expected to contribute to the literature especially in two aspects.

Firstly, scales measuring examinees' attitudes toward CAT are frequently used in health sector, recruitment processes and employee evaluation procedures. In fact, recent research (e.g., Martin *et al.*, 1983; Martin *et al.*, 1983; Moreno *et al.*, 1984) has been conducted largely in a military environment. In contrast, scales measuring examinees' attitudes toward CAT in the field of education is limited (Chuesathuchon, 2008). Therefore, investigating attitudes toward CAT in the field of education is valuable to improve CAT for future educational testing.

Secondly, the scale developed in the current study is precious because the scale was developed by collecting data from examinees after an actual CAT experience. Reviewing the literature showed that previous research relied on data collected from examinees without experiencing a CAT. On those research studies, examinees were mostly informed about how CAT proceeds theoretically and expected to respond questionnaires or someone observed test-takers during CAT experience and filled out obser-

vation schedule simultaneously. Moving from theory to practice is important to better shape the CAT process. Therefore, developing a scale measuring attitudes toward CAT by collecting data from examinees just after they have real experience on CAT was expected to make great contribution to the literature.

1.3. Purpose of the Study

The purpose of this study is to develop a scale to measure examinees' attitudes toward CAT. The reliability and validity evidence of the scores obtained through the scale were provided.

1.4. Research Questions

The research questions of the study:

(i) To what extent are the scale scores used to assess test takers' attitudes towards CAT are reliable?

(ii) To what extent are the scale scores used to assess test takers' attitudes towards CAT valid?

(iii) To what extent do the scale scores used to assess test takers' attitudes towards CAT exhibit measurement invariance with respect to gender groups?

2. LITERATURE REVIEW

2.1. Attitudes toward CBA

CBA implementations in educational testing have grown up gradually (Ghaderi *et al.*, 2014; Feketéné *et al.*, 2002; Organization for Economic Cooperation and Development, 2010). Investigating examinees' attitudes toward CBA is valuable for the future implementations be more successful (Lilley *et al.*, 2011; Özbaşı, 2016). Reviewing the literature demonstrated that there are certain studies that examined test-takers' attitudes and some related concepts such as experiences, acceptance, perceptions and reactions toward CBA (Klella, 2020; Nugroho *et al.*, 2018; OECD, 2010; Ogilvie *et al.*, 1999; Stricker *et al.*, 2004; Şanlı, 2003; Zilles *et al.*, 2015). An overview of certain studies is provided in Table 2.1.

Regarding the CBA, almost all the studies revealed that examinees' attitudes toward CBA are overwhelmingly positive (Ogilvie, Trusk *et al.* 1999). Investigated the attitudes of 202 medical students who attended both CBA and P and P test in a first-year cell biology and histology course. Furthermore, the impact of CBA on students' study habits was observed. The findings of this study revealed that, compared to P and P test, the majority of students found CBA more entertaining, time-efficient and informative. The authors also observed that study habits were impacted positively by computerized administration. Even, because of favorable feedback received from computer examination, final and mid-course exams were administered in CBA (Klella, 2020). Recently conducted a study to fundamentally examine students' reactions and teachers' attitudes toward CBA in English class. Observation checklist was used to identify students' reactions while they were experiencing CBA and concluded that students found computers interesting and enjoyed CBA. To find out teachers' attitudes toward CBA, a questionnaire was used, and found that attitudes of teachers toward CBA are positive. On the other hand, teachers had worries for applying CBA in practice due to not having adequate laboratories. Those teachers also highlighted one

crucial point that instructors require intensive courses covering computer skills and essentials (Stricker *et al.*, 2004). Investigated test takers' acceptance CBA version of TOEFL. In fact, relations between acceptance, test performance, general attitude about test and other potential determinants such as computer familiarity and total test anxiety were examined. The central finding of this extensive study was that test takers' attitudes toward CBA were moderately positive. The case study of (Şanlı, 2003) aimed to gain students' perceptions about online assessment. An exam website has been developed and administered to participants consisting of 46 third-year students from the Department of Computer Education. As with prior research, data collected via the questionnaire and user interviews proved that students had positive attitudes toward CBA.

At present, however very little work has attempted to compare examinees' attitudes toward P and P tests and CBA, what has become clear is that most test-takers preferred CBA over P and P test. Prominent example of such research is conducted by (Zilles *et al.*, 2015). In this study, experiences of 200 students through class exams in CBA were described and revealed that 75% of students prefer CBA, 12% traditional written exams and 13% had no preference. The experimental study of (Nugroho *et al.*, 2018) where one of the classes in a public university took P and P test as a science final exam while the other took CBA in the same science subject, aimed to examine the difference between two groups in terms of perception about the administration. It revealed an extremely interesting finding that students preferred to take test with different test method other than what they were taking.

The study of (Organization for Economic Cooperation and Development, 2010) presented the most comprehensive overview by investigating significant aspects of computer-based version of science in PISA in comparison to P and P in three countries of Korea, Iceland and Denmark. Consistent with the findings of (Zilles *et al.*, 2015), it was revealed that most students preferred to do CBA than P and P. Moreover, in the same study, it was found that motivation of test-takers in CBA was higher than in P and P in all three countries and students enjoyed more in CBA. Additionally, a

gender-based comparison was also made and it was suggested that females had much less confidence than males in most ICT activities, similar to the results of ICT PISA 2003.

Table 2.1. Studies of examinees' attitudes toward CBA.

Study Reference	**Findings**
Ogilvie, Trusk and Blue, 1999	• Students' attitudes toward the computer assessment in the course were very positive.
Şanlı, 2003	• Students had positive attitudes toward online assessment.
Stricker, Wilder and Rock, 2004	• Test-takers' attitudes toward computer-based TOEFL appeared to be moderately positive.
OECD, 2010	• Students enjoyed the CBT more than P and P. • Most students preferred to take test in computerized version rather than P and P. • Males were more confident than females in most ICT activities.
Zilles, Deloatch, Bailey, Khattar, Fagen-Ulmschneider, Heeren, Mussulman and West, 2015	• 75% of students preferred computerized testing. • 12% of students preferred traditional written exams. • 13% of students had no preference.
Nugroho, Kusumawati and Ambarwati, 2018	• Significant difference between two groups of students on their perception on CBT and PPT.
Klella, 2020	• Students found computers interesting and they enjoyed the testing process. • Teachers had positive attitude to use of computers in testing.

2.2. Attitudes toward CAT

CAT is a new form of implementing a test through a computer that adapts to the test taker's ability level (Weiss, 1985). Reviewing the literature demonstrated that there are certain studies examined test-takers' attitudes and some related concepts such as experiences, acceptance, perceptions and reactions toward CBA.

There is only a limited amount of work specifically focusing on attitudes of examinees toward CAT. Findings of most of those studies showed that attitudes of examinees toward CAT are largely positive (Baghi *et al.*, 1991; Legg *et al.*, 1992; Moe *et al.*, 1988; Schmidt *et al.*, 1978; Vicino *et al.*, 1997). Meanwhile, CAT raised many questions regarding its characteristics that examinees like or dislike (Vicino *et al.*, 1997). Suggested that examinees no more faced constraints in reading on the screen compared to paper. Examinees like CAT features of shorter length and the ability to receive feedback (Tonidandel *et al.*, 2000; Vispoel *et al.*, 1994; English *et al.*, 1977) confirmed that tailored testing required 75% less time and examinees liked shorter tests. Furthermore, it was suggested that examinees were less or no longer anxious with the CAT than with P and P test. Girls found P and P test more worrisome than boys due to time limit in P and P test (Özbaşı, 2016). Collected opinions of 205 students who experienced P and P and CAT and found that students had higher degrees of motivation in CAT (Daphine *et al.*, 2020) have focused on attitude of students towards an online CAT in physics education. The teacher observed and filled out the observation schedule while the students were taking CAT. Attitudes towards CAT overlap with the previously discussed studies. Key finding indicated that CAT implementations improve the quality of physics education, so academicians and policy makers might collaborate to expand the use of CAT applications. In the studies of (Özbaşı, 2016) and (English *et al.*, 1977) have bridged CAT characteristics and achievement scores, then obtained exact conflicting evidence (Özbaşı, 2016). Suggested that CAT characteristics had an impact on achievement scores (English *et al.*. 1977). Refutes this view as they found there is no difference between test scores of examinees experience CAT and P and P test.

Some negative characteristics of CAT were also reported (Legg *et al.*, 1992) found that examinees were bothered by the inability to review or skip items (Baghi *et al.* 1992; Vicino *et al.*, 1997; Tonidandel *et al.*, 2000; were in line with Legg *et al.*, 1992) in that not being able to review or modify answers to previous items affect examinees' attitudes toward CAT negatively (Baghi *et al.*, 1992). Found a great problem that many examinees faced in their study of CAT use, that is, scrolling through the reading paragraphs. Despite negative attitudes toward some characteristics of CAT, as a result of reviewing literature, it can be concluded that examinees preferred CAT over P and P test (Baghi *et al.*, 1991; Vicino *et al.*, 1997).

Table 2.2. Studies of examinees' attitudes toward CAT.

Study Reference	**Findings**
English, Reckase and Patience, 1977	• Tailored tests were found less stressful. • 75% time savings were found in CAT.
Schmidt, Urry and Gugel, 1978	• Overall attitudes are overwhelmingly positive toward CAT. • To be not able to go back and review answer is mentioned as the worst thing. • Requiring less time is mentioned as the best thing.
Moe and Johnson, 1988	• Overall reactions to the computerized test were overwhelmingly positive.
Baghi, Ferrara and Gabrys, 1991	• Test-takers preferred the CAT over P and P test.
Baghi, Ferrara and Gabrys, 1992	• Scrolling problems through reading paragraphs were recorded.
Legg and Buhr, 1992	• Attitudes of examinees toward computerized testing are generally very positive for all the groups studied. • Reaction to the inability to review questions after a response had been entered is the least positive one.

Table 2.2. Studies of examinees' attitudes toward CAT. (cont.)

Study Reference	**Findings**
Vispoel, Rocklin and Wang, 1994	• Test-takers disliked the inability to review or skip items. • Test-takers liked the shorter length of CAT and ability to receive feedback.
Vicino and Moreno (1997)	• Examinees preferred the CAT over P and P test. • Test-takers disliked not being able to review or change responses to previous question.
Tonidandel and Quinones, 2000	• Certain features of CAT may adversely affect examinees' reactions.
Özbaşı, 2016	• CAT applications increase students' motivation. • Girls found P and P tests more worrisome than CBA.
Daphine, Sivakumar, Selvakumar, 2020	• CAT implementations improve quality of physics education.

2.3. Scale Development Studies Measuring Attitudes toward CBA/CAT

Only a very few studies aimed to develop a scale measuring test-takers' attitudes and other related concepts such as experiences, perceptions, acceptance and reactions to CBA and CAT. In fact, CAT-related scale developments studies are relatively less than CBA-related ones. The common point of these studies is that constructs in measurement models had their foundation in solely literature, not based on any theoretical model.

Table 2.3. Scale development studies measuring.

Study Reference	**Constructs**	**Sample Items**	**CBA/ CAT**
Burke, Normand and Raju, 1987	General Acceptance of CBA	.	CBA
	Ease of Taking CBA	.	
Smith and Caputi, 2004	Ease of Use	I felt more comfortable completing the test on paper than computer.	CBA
	CBT-Confidence	Computerized tests require too much experience with computers.	
Chuesathuchon, 2008	Like and Interest in CAT	I am happy and enjoyed doing a computerized adaptive test.	CAT
	Confidence and Use of CAT	The computerized adaptive testing makes me want to study Mathematics.	
	CAT as Modern and Useful CAT as Reliable, Fair and Good CAT Recommendations	Computerized adaptive testing saves money	
	CAT as Reliable, Fair and Good CAT Recommendations	Computerized adaptive testing is fair for all students.	
	CAT Recommendations	I wish I could take a computerized adaptive test in a Mathematics test competition.	

Table 2.3. Scale development studies measuring. (cont.)

Study Reference	**Constructs**	**Sample Items**	**CBA/ CAT**
Lim, 2020	Ease of Use	I knew where to provide my response(s) to each question in the e-exam.	CBA
	Involvement	I could concentrate while doing the e-exam.	
	Experience	I like the e-exam experience.	

Burke *et al.* (1987) and (Smith *et al.*, 2004) focused on the attitude of the examinees toward CBA. In the former, a short attitude questionnaire was administered after the completion of 1/2 hour computer-based ability test. A factor analysis of the attitudinal items produced two factors labeled as general acceptance of CBA and ease of taking a CBA. In the latter the psychometric properties of Attitude Towards CBA Scale were examined. The results of the explanatory factor analysis (EFA) yielded in a two-factor model as the ease of use and CBT-Confidence. Unlike (Burke *et al.*, 1987) and (Smith *et al.*, 2004; Lim 2020) focused on the concept of acceptance and aimed to develop and validate the CBA Acceptance Questionnaire for secondary and high school students in Singapore. Exploratory factor analysis and confirmatory factor analysis yielded in a three-factor model as the ease of use, involvement, and experience with a reduction from 21 items to 13 items.

Unlike the earlier research, (Chuesathuchon, 2008) attempted to develop a scale measuring attitudes toward CAT. IRT was used to calibrate the test items. Constructs were based on in-depth literature review and attitude was conceptualized from five aspects as like and interest in CAT, confidence with and use of CAT, CAT as modern and useful, CAT is reliable, and CAT recommendations.

2.4. Prior Studies of TAM in CBA

Technology Acceptance Model is mostly used in the field of educational research specifically for e-learning, web-based assessment and computer-based assessments in order to investigate how users come to accept or reject and use a technology (Davis *et al.*, 1989).

Regarding the literature, there is not any CAT-related research based on TAM in educational context while there is one in the health sector (Nikolaus *et al.*, 2014). However, there are some prior studies of TAM in CBA. In thesis study, (Alkış, 2010) identified factors that impact higher education students' perceptions of CBA for learning, using TAM as a conceptual framework. The data was obtained from 332 Middle East Technical University students and analyzed via confirmatory factor analysis, through structural equation modeling. It was yielded in five constructs as perceived ease of use, perceived usefulness, intention, computer attitude and anxiety. The result indicated that perceived usefulness is the biggest determinant of students' willingness to use CBA. Subsequent research, (Terzis *et al.*, 2011), built a model that illustrates the constructs that impact students' behavioral intention to use CBA. The proposed model is based on Technology Acceptance Model in addition to other previous models of Theory of Planned Behavior (TPB) and The Unified Theory of Acceptance and Usage of Technology (UTAUT). Perceived usefulness and perceived ease of use from TAM, and social influence, perceived playfulness, computer self-efficacy, self-management of learning, learning goal orientation and facilitating conditions from TPB and UTAUT were used in the measurement model. Additionally, personal outcome expectations and content were integrated in proposed model. According to results of this study, while perceived playfulness and perceived ease of use have a direct effect on use of CBA; perceived usefulness, social influence, computer self-efficacy, facilitating conditions, content and goal expectancy have solely indirect effects. These eight variables explain approximately 50% of the variance of behavioral intention (Terzis *et al.*, 2011) also investigated gender differences in perceptions and acceptance about computer-based assessment. Computer Based Assessment Acceptance Model (CBAAM) (Terzis

et al., 2011) was extended to take gender into consideration. Constructs which impact male and female students' behavioral intention to use CBA were identified. The questionnaire was completed by 117 female and 56 male students. The results indicated that female students are more likely to use CBA if it is easy to use. On the other side, male students are more likely to use CBA if it is playful. In other study, (Dizon, 2016), it was aimed to measure Japanese EFL students' perceptions of Internet-Based Tests with three variables of TAM as perceived usefulness, perceived ease of use, behavioral intention to use. It was highlighted that TAM was taken as the base model because it is the most effective and widely employed model to predict a user's adoption of new technologies (Lee *et al.*, 2003). 80 Japanese university students responded to the adapted TAM questionnaire. A central finding was that three variables were positively and highly correlated with each other.

3. METHODOLOGY

3.1. Sampling/Participants

When developing a scale, sufficient amount of sample is needed to provide reliability and validity evidence of the scores. Sample size should be sufficiently large to eliminate subject variance and obtain more accurate results. The common rule of thumb for sample size is that there should be at least 10 participants for each item, making an ideal of 15:1 or 20:1 (Clark *et al.*, 1995; DeVellis, 2003).

There are two primary categories of sampling as random sampling and non-random sampling. Random sampling refers to every member of the population having equal and independent probability of being selected. In contrast, in non-random sampling, sample selection is based on some criteria such as availability of examinees, economic conditions and judgments of researchers (Fraenkel *et al.*, 2003). This study was conducted in two phases: pilot study and final study. For both phases, the sample was selected by convenient sampling, which is one of the non-random sampling methods. Determining the appropriate sample size, suggestions of (Clark *et al.*, 1995) and (DeVellis, 2003), 20 participants for each item, were taken into consideration. Schools that agreed to take part in the study were included. The distribution of sample for pilot and final studies according to school type and gender are presented in Table 3.1.

In the pilot study, the initial form of the scale was administered to 311 fourth-grade primary school students in May and June 2023. Data was collected from 5 private schools located in Denizli, Turkey. The collection of sample data was limited to private schools due to lack of laboratories in public primary schools. Students were started with a live CAT that were developed for 4th grade mathematics topics measuring knowing, applying and reasoning abilities of students. Just after the CAT, the Examinees' Attitudes Toward Computerized Adaptive Test Scale was administered to the students.

In the final study, the revised scale was implemented to 241 fifth-grade middle school students in September and October 2023. Data was collected from 2 public and 3 private schools located in Denizli and Aydın, Turkey different from the pilot study.

Table 3.1. The fit indexes.

Phases	**School Type**	**Gender**	***n***	**Total**
Pilot Study	Private	Boys	147	311
		Girls	164	
	Public	Boys	-	0
		Girls	-	
Final Study	Private	Boys	24	40
		Girls	105	
	Public	Boys	105	201
		Girls	96	

3.2. Instrument

3.2.1. Test Construction Process

According to (DeVellis, 2017), there are eight steps involved in developing a scale (Crocker *et al.*, 2008), on the other hand, recommend ten steps (1986). This section explains how to develop a scale based on (DeVellis, 2017; Crocker *et al.*, 2008).

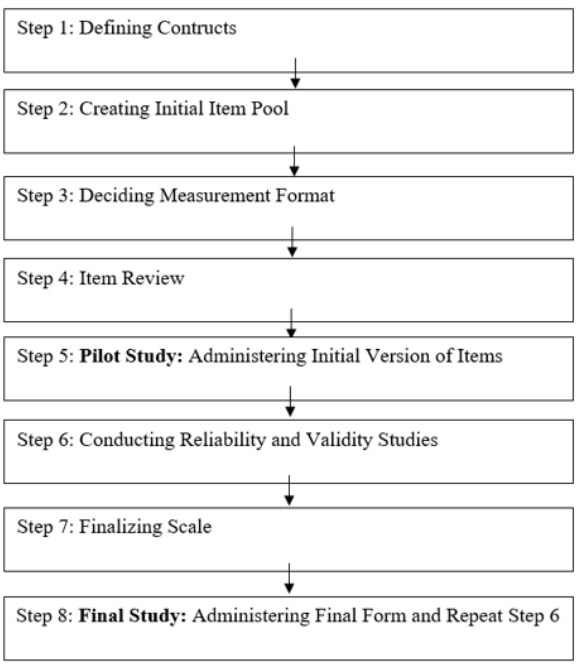

Figure 3.1. Scale development steps.

3.2.1.1. Defining Constructs.. Definitions and scopes of constructs should be clearly defined so that the researcher can develop a scale what he or she genuinely intends to do. In defining constructs, adhering to at least one theoretical model benefits the researcher. Furthermore, scopes of dimensions can be relatively broad or narrow with regard to the situations to which scales apply and the constructs they cover. This entire process is fundamentally shaped depending on the goals of the researcher (Crocker *et al.*, 2008; DeVellis, 2017). After defining the construct, preparing the table of specification is a necessary and essential part for content validation (Crocker *et al.*, 2008) in order to define the content domains being measured and to guarantee that a fair and representative sample of items appears on the questionnaire or test. Table of specification is comprised of the categories, example items in these categories and the item numbers.

3.2.1.2. Creating Initial Item Pool.. The first step is to generate a large pool of items after articulating the goal of the scale clearly. The content of each item should principally reflect the latent trait underlying them. Characteristics of good items can be listed as follows (Crocker *et al.*, 2008; DeVellis, 2017):

- Avoiding exceptionally lengthy items,
- Avoiding unnecessary wordiness,
- Being composed of simple sentences rather than complex or compound ones,
- Avoiding double-barreled items,
- Being suitable for the level of readability,
- Avoiding offensiveness and appearance of bias,
- Comprising vocabulary that respondents can easily understand,
- Avoiding grammatical errors (e.g., spelling errors, etc.),
- Avoiding statements that can have more than one interpretation.

3.2.1.3. Deciding Measurement Format.. In this step, the purpose is to decide the format of the items. There are numerous formats to present items and response options.

Most common types are Thurstone Scaling, Guttman Scaling, Likert Scale and Semantic Differential. Likert scale is frequently used in instruments measuring opinions, beliefs and attitudes (DeVellis, 2017). Some specific guidelines that could be useful when writing in Likert format are as follows (Crocker *et al.*, 2008):

(i) Use present tense for statements or questions,
(ii) Attempt to have almost equal amounts of statements reflecting both positive and negative emotions,
(iii) Choose short statements, rarely more than 20 words,
(iv) Avoid using universals like all, always, none and never which frequently cause ambiguity in statements,
(v) Avoid using indefinite qualifiers such as merely, seldom, only, many, few or just,
(vi) Avoid using "if" or "because" clauses in statements,
(vii) Avert the use of negatives (e.g., never, none, not).

In Likert-type, the respondents read each statement and then, to indicate the degree of strength of their endorsements, respondents select a response from a four, five, seven and etc.-point continuum ranging from "Strongly disagree" to "Strongly agree". While scoring items with a graded response continuum, 1 point is given to the response showing the lowest amount of positive endorsement toward the construct; 2 points are received to the response displaying the next highest degree and so forth (Crocker *et al.*, 2008).

3.2.1.4. Item Review.. Subject matter experts review the item pool before administering the scale to improve the content validity (DeVellis, 2017). First of all, subject matter experts are asked to rate how relevant they think each item is to what researchers intends to measure. It should be definitely highlighted that the construct and the content of item should be relevant to each other conceptually. Even when the construct and the content of item is relevant to each other, there may be a problem in wording. For this reason, reviewers also evaluate the items' clarity and conciseness. Finally, experts point out ways of tapping the phenomenon that researchers have failed

to comprise. It is crucial that final decision to accept or reject the advice of subject matters is left to the scale developer after carefully considering all the suggestions (DeVellis, 2017).

3.2.1.5. Pilot Study: Administering Initial Version of Items.. It is recommended for scale developers to test the items prior to proceeding with final study. The results of pilot study provide extensive information regarding which items should remain in the scale and which should be eliminated or be revised (Crocker *et al.*, 2008).

3.2.1.6. Conducting Reliability and Validity Studies.. Reliability and validity evidence of scale scores should be provided. A comprehensive explanation of how reliability and validity studies are carried out is included in the 3.3 Data Analysis section.

3.2.1.7. Finalizing Scale.. The test developer makes final decisions about which items to retain, eliminate or revise on the basis of implementation on large enough sample and in-depth analysis of the scale's initial version. At that point, scale can be finalized (Crocker *et al.*, 2008).

3.2.1.8. Administering Final Form and Repeat Step 6.. The finalized version of scale should be applied to participants, taking into size and composition of development sample. Reliability and validity studies should be repeated with data collected from final study.

3.2.2. Development of Scale on Examinees' Attitudes toward CAT

This section explains how to integrate test construction steps and suggested practices in current study to develop scale measuring examinees' attitudes toward CAT.

3.2.2.1. Defining Constructs.. Technology Acceptance Model presented by (Davis, 1989) provided a major conceptual framework for the current study. There are three

dimensions in the current study: perceived usefulness, perceived ease of use and perceived enjoyment. Perceived usefulness and perceived ease of use were rooted in TAM whereas perceived enjoyment had its foundation in literature. Some studies concluded that enjoyment has a significant effect on intentions to use information technology (Davis *et al.*, 1992; Teo *et al.*, 2011). In fact, there were a number of models which combine TAM dimensions and perceived enjoyment (e.g., Karagöz, 2022; Lee *et al.*, 2005; Martínez-Torres *et al.*, 2008; Mun *et al.*, 2003; Saadé *et al.*, 2005).

TAM designates the relationships between system design features, perceived usefulness, perceived ease of use, attitude toward using and actual usage behavior (Davis, 1993). The relations between TAM's variables are shown in Figure 3.2 below.

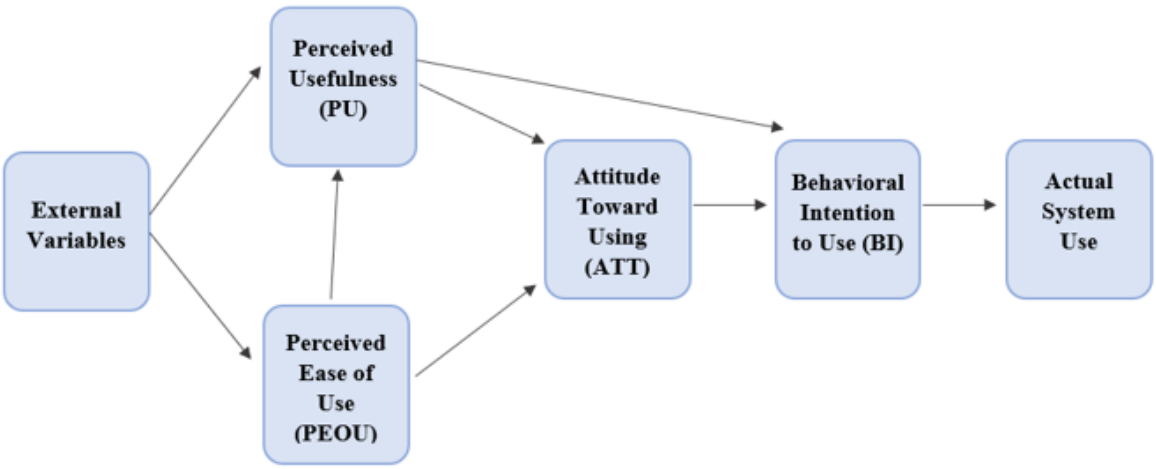

Figure 3.2. Technology Acceptance Model (TAM) (Davis, 1993).

Perceived usefulness is defined as "the degree to which an individual believes that using a particular system would enhance his or her job performance" (Davis, 1989, p.320). In the current study, perceived usefulness was defined as the degree to which an examinee believes that CAT is advantageous to traditional paper-pencil test. One of the items is "The test I just answered took less time than my other exams". Examinees who think that they are able to complete the test quicker in CAT are expected to have higher grades from perceived usefulness part of the questionnaire. Higher grades from this part mean higher degree of taking more advantage of CAT rather than paper-pencil test and becoming inclined to develop positive attitudes toward CAT.

Perceived ease of use is defined as "the degree to which an individual believes that using a particular system would be free of physical and mental effort" (Davis, 1989, p.320). In the current study, perceived ease of use was defined as the degree of an examinees' beliefs of using CAT being free of difficulty or effort. Within this dimension, it is examined whether or not CAT system can be easily used by the examinees with lower level of mental and physical exhaustion. The example item included is "Reading the questions on the screen was challenging for me". Examinees who get high scores from this part are expected to find CAT system difficult to use and to tend to develop negative attitudes toward using CAT.

Perceived enjoyment is defined as the extent to which the activity of using technology is regarded as enjoyable in its own right without any consideration to performance outcomes that may be anticipated (Carroll *et al.*, 1988; Deci, 1971; Venkatesh, 2000). Within this dimension, the participants' overall satisfaction with the CAT system is examined. Both negative and positive statements are reflected in this dimension. Getting higher scores from this item that "The test was enjoyable". Means positive attitudes toward using CAT. In contrast, to get higher scores from the item that "Instead of taking the test just before, I would like to do something else." means negative attitudes toward CAT.

3.2.2.2. Creating Initial Item Pool.. In the current study, perceived usefulness, perceived ease of use and perceived enjoyment were thoroughly defined, then the scale was prepared in Turkish. Within the framework of clear and detailed descriptions of dimensions, items were written as follows:

Items of Perceived Usefulness:

- Biraz önce yanıtladığım testte, öğretmenimin yaptığı testlere göre daha az soru çözdüm. (I solved less questions in the test I just answered compared to test my teacher gave.)

- Biraz önce yanıtladığım test, öğretmenimin yaptığı testlere göre daha hızlı bitti. (The test I just answered was completed faster than the tests my teacher gave.)
- Benim bilgi seviyeme uygun sorularla karşılaştım. (I encountered the questions that were appropriate to my level of knowledge.)
- Soruların boş bırakamadığım için rastgele bir şık işaretlediğim oldu. (Due to inability to leave questions, I marked a choice randomly.)
- Biraz önce yanıtladığım testteki sorular, öğretmenimin yaptığı testlerdeki sorulardan daha kolaydı. (The questions on the test I just answered were easier than the questions on the tests my teacher gave.)

Items of Perceived Ease of Use:

- Ekranda işaretleme yapmak yorucuydu. (Marking on the screen was exhausting.)
- Bilgisayar kullanımı ile ilgili öğretmenimden yardım istedim. (I requested assistance from my teacher about computer use.)
- Soruları ekrandan okumakta zorlandım. (Reading the questions on the screen was challenging for me.)
- Bilgisayardan okumak zoruları anlamamı zorlaştırdı. (Reading on the computer complicated comprehension of the questions.)
- Soruları bilgisayardan okuyup, işlemleri kağıtta yapmak beni yordu. (Reading the questions on the computer while making calculations on paper made me tired.)

Items of Perceived Enjoyment:

- Eğlenceli bir testti. (The test was enjoyable.)
- Testi çözmeyi sevdim. (I liked taking the test.)
- Biraz önceki testi çözmek yerine, başka bir şey yapmak isterdim. (Instead of taking the test just before, I would like to do something else.)
- Testin hemen bitmesini istedim. (I wanted the test to end immediately.)
- Testi çözmek zevkliydi. (Taking the test was pleasurable.)

3.2.2.3. Item Review.. In the current study, content validity was checked by using item-objective congruence. The entire item review process is described below Validity Evidence for the Current Study of 3.3.2.1 Content Validity section.

3.2.2.4. Pilot Study: Administering Initial Version of Items.. Following item review, initial version of the scale was developed. In the pilot study, students responded the initial version of the scale with 15 items after experiencing a live CAT test on mathematics.

3.2.2.5. Conducting Reliability and Validity Studies.. The methods used to provide reliability and validity evidence in current study are described comprehensively in 3.3 Data Analysis section.

3.2.2.6. Reviewing Items and Finalizing Scale.. According to item fit indices, factor loadings, IRT item parameters and item characteristic curves; items were evaluated, and final form of the scale was decided.

3.2.2.7. Administering Final Form and Repeat Step 6.. The final form of the scale was administered and procedures in step 6 were followed for reliability and validity evidence of the test scores.

3.3. Data Analysis

3.3.1. Reliability

Reliability refers to the overall consistency of the scores obtained (DeVellis, 2017). A reliable scale carried out on similar groups of respondents under similar circumstances would yield similar results (Cohen *et al.*, 2018). Test scores are usually not identical because of a variety of factors; such as, anxiety and different testing conditions. Such factors cause errors in measurement. However, results should be close to each other

even if they are not exactly the same. As a specific example to make it clear; if the test is reliable, a student who receives a low score from this test on the first time should also receive a low score on the subsequent administration of the test (Fraenkel *et al.*, 2003). Following are three main types of reliability: stability, equivalence and internal consistency (Carmines *et al.*, 1979).

Reliability as stability, usually called test-retest, is a measure of consistency over time. A test is administered to a group of examinees once, then repeated to the same participants. The coefficient of stability between the two sets of scores, the correlation coefficient obtained from the test-retest procedure, is computed (Cohen, Manion *et al.*, 2018). The time that should elapse between two administrations of the test is critical. The time period should be long enough to prevent participants from remembering the initial administration but not so long as to allow a change in situational factors (Cooper *et al.*, 2001). The second way to achieve reliability is inter-rater reliability. Inter-rater reliability refers to the degree of agreement between multiple raters. Inconsistencies between assessment decisions may emerge because human judgment is fallible. Inter-rater reliability can be achieved by ensuring that all raters provide similar decisions (Cohen *et al.*, 2018).

Finally, internal consistency is defined as a measure based on the correlations between test items (DeVellis, 2017). In contrast to the fact that test-retest method and alternative forms method require the test to be administered twice, internal consistency methods need only a single administration of the instrument (Cohen *et al.*, 2018; Fraenkel *et al.*, 2003). Split-half technique and alpha coefficient are two primary methods of estimating internal consistency (Cohen *et al.*, 2018). Within split-half procedure, the test items are divided into two halves. For instance, one half may be composed of even numbers while the other half is composed of odd numbers. Both halves are administered to the same participants and graded. If the test exhibits split-half reliability, the scores obtained from each half should have a strong correlation with one another.

The alpha coefficient, often known as Cronbach's Alpha, is another measurement of reliability as internal consistency. (Cohen *et al.*, 2018). For the Cronbach's Alpha coefficient, the following guidelines can be used (Cichetti, 1994):

Table 3.2. The cronbach's alpha criterion.

Coefficient of Cronbach's Alpha	**Internal Consistency**
$a > 0.9$	Excellent
$0.9 > a \geq 0.8$	Good
$0.8 > a \geq 0.7$	Acceptable
$0.7 > a \geq 0.6$	Questionable
$0.6 > a \geq 0.5$	Poor
$a < 0.5$	Unacceptable

McDonald's Omega is one of the suggested alternatives to Cronbach's Alpha (McNeish, 2018). Omega is more powerful than Alpha against deviations from assumptions of alpha and generally found as more appropriate measure of internal consistency (Stensen *et al.*, 2022). The difference between Alpha and Omega values can be important, but will not big. Differences in values are rooted in the deviations from assumptions (Yang *et al.*, 2011). Similar to Alpha, omega value is often regarded as acceptable if the estimate is 0.70 or above (McNeish, 2018). In congeneric tests, the reliability coefficients might be lower than actual values (Lucke, 2005, 2005; Zinbarg *et al.*, 2005). Reliability Evaluation in the Current Study.

Cronbach's Alpha and McDonald's Omega were used to provide evidence for reliability of the test scores in the current study. They are most frequently used to test reliability for scales consisting of several Likert-type items (Gliem *et al.*, 2003; Laerd Statistics, 2013). The values were interpreted according to (Cichetti, 1994) criteria presented above. Both coefficients were calculated using SPSS 29 to provide evidence for reliability of test scores.

3.3.2. Validity

Validity fundamentally concerns coming to the right conclusions based on the information obtained from an assessment (Fraenkel *et al.*, 2003). It might be defined as the extent to which a test measures what it is actually intended to measure (DeVellis, 2017). In research, reliability is a required precondition while being an insufficient condition for validity. Although a high reliability coefficient suggests that test takers' results are consistent, it does not guarantee that the examiner's inferences are tenable (Cohen *et al.*, 2018). According to (Cronbach, 1971), validation is the procedure of gathering and evaluating evidence to support such inferences. There are essentially three types of validity; (1) Content-related validity, (2) Criterion-related validity and (3) Construct-related validity (Crocker *et al.*, 2008; DeVellis, 2017; Fraenkel *et al.*, 2003).

3.3.2.1. Content Validity.. Firstly, the extent to which a specific set of items accurately covers a content domain, item sampling adequacy, is known as content validity. The instrument format is the main point in content validity. A few aspects of "format" are clarity of printing, type size and appropriateness of wording. Briefly stated, the content and format must match the sample of measured subject and descriptions of variables for high degree of content validity (Crocker *et al.*, 2008; Fraenkel *et al.*, 2003).

In content validation, creating and evaluating a table of specifications is crucial (Crocker *et al.*, 2008). Dimensions, examples of items in the dimensions and item numbers are involved in table of specifications. The aim of a table of specifications is to specify the measured content domains and to guarantee that a fair and representative sample of items included in the questionnaire/test. Then, experts are requested to examine the content and the instrument's format and investigate whether the items adequately sample the domain of interest following the development of the instrument' initial version. This procedure might be called "expert opinion". Steps can be summarized as follows (Crocker *et al.*, 2008; Fraenkel *et al.*, 2003):

(i) Content domain is defined clearly by the researcher.

(ii) Definitions of what he/she wants to measure, intended sample and instruments are given to subject matter experts. In addition, structured framework for the task of matching items to the content domain is provided to the experts.

(iii) Experts evaluate the items by reading them, then rating how well items math to a specified objective.

(iv) The results are collected and summarized via matching procedures.

(v) The researcher rewrites the items if necessary and resubmits to the experts.

For summarizing the results, in stage 4, index of item-objective congruence is calculated as below (Hambleton, 1980; Rovinelli *et al.*, 1977). The index of congruence of item to objective can be calculated by the formula

$$l_{ik}\frac{N}{2N-2}(\mu_k - \mu), \tag{3.1}$$

where $N$is the number of objectives, μ_k is the judges' mean rating of item i on the k th objective, and μ is the judges' mean rating of item i on all objectives.

According to (Rovinelli *et al.*, 1977), the value of index should be at least 0.50. Higher values mean congruent with the factor; while lower ones mean that items should be checked.

3.3.2.2. Criterion Validity.. The relationship between scores acquired using the particular instrument and scores obtained from another external criterion is known as criterion-related validity (Cohen *et al.*, 2018; Fraenkel *et al.*, 2003). A criterion is a second test or another assessment method that is supposed to measure the same variable. A criterion-related validation study generally follows the steps below (Cohen *et al.*, 2018):

(i) Defining a criterion behavior and a method to measure it.

(ii) Defining a representative sample to administer a test.

(iii) Applying the test and saving rating scores.

(iv) Getting a measure of a performance on the criterion for each test-taker once the criterion data are available.

(v) Detecting the degree of relationship between scores of test and criterion performance.

There are two fundamental forms within this kind of validity: predictive validity and concurrent. Predictive validity is defined as how well the test results anticipate criterion measurements in the future. The key question is whether or not the test results successfully predict outcomes. If there is a high correlation between the results of two administrations, predictive validity is achieved (Cohen *et al.*, 2018). On the other hand, concurrent validity describes a correlation between the test scores and criterion measurement performed at the time the test was administered. (Crocker *et al.*, 2008). If tests produce close or identical results; concurrent validity is achieved. (Fraenkel *et al.*, 2003).

3.3.2.3. Construct Validity.. Construct validity is directly related to how a variable is related to other variables theoretically (Cronbach *et al.*, 1955). It is described as the ability of a measurement tool to represent a concept that is not directly observable such as "intelligence" and "creativity". These abstract concepts should be operationally defined.

There are four frequently used methods for construct validation: (1) Correlations between a measure of the construct and designated, (2) Differentiation between groups, (3) The multitrait-multimethod matrix and (4) Factor analysis (Crocker *et al.*, 2008).

Firstly, to explain correlations between a measure of the construct (eg. school performance) and designated (eg. intelligence test), it is necessary to consider correlational evidence of the relationship between construct and designated. School performance and intelligence tests might not be identical constructs. However, there should be a relationship between them. Otherwise, the construct of intelligence would be less useful (Crocker *et al.*, 2008).

Secondly, an example of differentiation between groups is comparing the mean scores of the subjects who received a specific treatment and no treatment on any construct to see whether or not two groups diverge in the hypothesized direction. Unable to find anticipated differences would cast doubts on the scores. Additionally, the adequacy of instrument would be questionable.

Thirdly, (Campbell *et al.*, 1959) developed the multitrait-multimethod matrix (MTMM) as a method for evaluating construct validity. Convergent validity coefficients and discriminant validity coefficients might be used to contrast how a measure is associated to other measures. Correlations between measures of the same construct obtained using several methods of assessment are known as convergent validity coefficients, while between different constructs are known as discriminant validity coefficients. Ideally, convergent validity coefficient should be high and discriminant validity coefficients should be very low.

Final procedure is factor analysis which enables researcher to investigate the instrument's factor structure (Crocker *et al.*, 2008; Fraenkel a*et al.*, 2003). Factor analysis is defined as a statistical approach that can be used to examine the interrelationship between many different variables and to explain these variables in terms of their shared underlying dimension called "factor" (Ebel *et al.*, 1991). There are two types of factor analysis as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Exploratory factor analysis (EFA) is a statistical technique for identifying the underlying structure of a sizable number of variables. Confirmatory factor analysis (CFA) aims to assess how well the collected data fits the proposed model. CFA has five steps. Guidelines for each step are as follows briefly (Ullman, 2006):

(i) Model Specification By using research in literature, which model to examine and validate should be specified. Consideration should be given for an inclusion of an important variable or for the removal of an unimportant variable from the model at this stage (Ullman, 2006).

(ii) Model Identification Based on the sample data that yield the sample covariance matrix and the theoretical model suggested by the population covariance matrix, a model is identified if each parameter in the model has a unique solution (Ullman, 2006). Parameters that are estimated should be more than data points (Ullman, 2001).

(iii) Model Estimation Several estimation methods exist to get estimates. Maximum Likelihood and Weighted Least Squares are frequently used ones. The former is for continuous observed data whereas the latter is for categorical data (Ullman, 2006). It should be noted that sample size is critical and 20 participants per observed variable is recommended (Clark *et al.*, 1995; DeVellis, 2003; Ullman, 2006).

(iv) Model Testing Once a model is specified, identified and parameters are estimated, next step is to determine whether or not this model is "good". A "good model" means there is a fit between population covariance matrix and sample covariance matrix. Indices such as basically Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI) and Tucker-Lewis index (TLI) are used for fit evaluation (Ullman, 2006). RMSEA has a range of 0 to 1. A good fitting model has values under 0.06 (Ullman, 2001). Values higher than 0.10 indicate a model that fits data poorly (Browne *et al.*, 1993). The most commonly recommended comparative fit indices are Tucker-Lewis index (TLI) and Comparative Fit Index (CFI), both of which compare independence model and estimated model using degrees of freedom (Ullman, 2001). CFI ranges from 0 up to 1 and TLI and CFI values over 0.95 indicate a good-fitting model (Hu *et al.*, 1999; Ullman, 2001).

(v) Model Modification If numerous fit indices are unsatisfactory, it is recommended to modify the model in order to enhance the fit. While modifying a model, the change should be justified by a theory and logical arguments (Ullman, 2006).

Validity Evidence for the Current Study To provide validity evidence for content validation, table of specification and expert opinion were done. As a first stage; table of specifications which involves dimensions, example items in dimensions and item

numbers were created and demonstrated below. As a result, alignment of the items and dimensions can be guaranteed (Frey, 2018). Secondly, "expert opinion" was followed. Basically, steps of "expert opinion" mentioned were followed. Experts was informed about the scope of the dimensions. A structured form consisting of the initial version of items and dimensions were given to 3 subject matter experts. For each item, experts rated whether the item is congruent with the given dimensions. It was rated +1 if the rater decided an item is congruent with dimension, 0 if the expert was not sure and -1 if the expert decided that the item is not congruent. Item-objective congruence index was calculated using the formula provided by (Hambleton, 1980; Rovinelli *et al.*, 1977). Items should be checked if values of indices were less than threshold. Experts were also be expected to share ideas about the format such as appropriateness of language. Ultimately, all the recommendations of experts about format and indices were evaluated with experts verbally in online meeting. If necessary, items were revised and rewritten. In the Section of 4.1.2.1 Content Validity Results, the traditional index of item-factor congruence values and revisions on items were provided comprehensively.

Table 3.3. Table of specification.

Table of Specifications			
Dimension Name	Item Numbers	Total Number	Percentages
Perceived Usefulness	2,7,11,13,15	5	33.3%
Perceived Ease of Use	1,5,8,9,12	5	33.3%
Perceived Enjoyment	3,4,6,10,14	5	33.3%
TOTAL		15	100%

Within construct-related validation, first of all, dimensions (perceived usefulness, perceived ease of use and perceived enjoyment) were operationally defined. Mplus was used to test factor structure of the scale via confirmatory factor analysis. Mplus is a statistical modeling application that provides researchers with a flexible instrument for data analysis (Muthén *et al.*, 2012). Fit indices in confirmatory factor analysis were checked. Values of TLI, CFI and RMSEA were summarized according to criteria mentioned above. The values of factor loadings were provided, and these were evaluated according to the fact that these values should be higher than 0.40. (Salkind, 2010).

3.3.3. Item Response Theory (IRT)

A fundamental objective of educational and psychological measurement is identifying the degree to which an individual possesses a latent trait such as reading and arithmetic abilities. To achieve this, it is vital to have a scale of measurement (Baker *et al.*, 2004). There are two common approaches used by psychometrics to understand how tests are analyzed and scored: Classical Test Theory (CTT) and Item Response Theory (IRT) (Crocker *et al.*, 2008). The main idea underlying CTT is that the raw test score of an examinee is the true score of the examinee with some degree of measurement error. CTT does not reveal how examinees with different degrees of ability levels perform on the item (Baker *et al.*, 2004; DeVellis, 2017). In contrast, under IRT, instead of focusing on the raw score, the main concern is whether an examinee answered each question correctly or not (Baker *et al.*, 2004). Moreover, IRT allows researchers to identify certain characteristics of items that are unaffected by who is answering them (DeVellis, 2017). It should be noted that IRT is a basic building block of CAT that has been gradually popular all around the world (Crocker *et al.*, 2008). Briefly stated, IRT is frequently promoted as a cutting-edge as an alternative to CTT and has attracted more attention lately because it occasionally offers more detailed information than CTT (Crocker *et al.*, 2008; De Boeck *et al.*, 2004; DeVellis, 2017; Embretson *et al.*, 2000; Nering *et al.*, 2010; Reise *et al.*, 2015). Basic concepts based on IRT will be presented in this section: item characteristic curve (ICC), item parameters, logistic models and assumptions respectively.

3.3.3.1. Item Characteristic Curve (ICC).. Item characteristic curve (ICC) is defined as "the nonlinear regression function of item score on the trait or ability measured by the test" (Hambleton *et al.*, 1984, p.25). At the lowest level of ability, the probability of the right answer is almost zero. As ability levels increase, the likelihood of getting the right answer gets closer to 1. This S-shaped curve is a typical characteristic curve demonstrated in Figure 3.3 (Baker, 2001). For polytomous items, ICC has the following shape (see Figure 3.4).

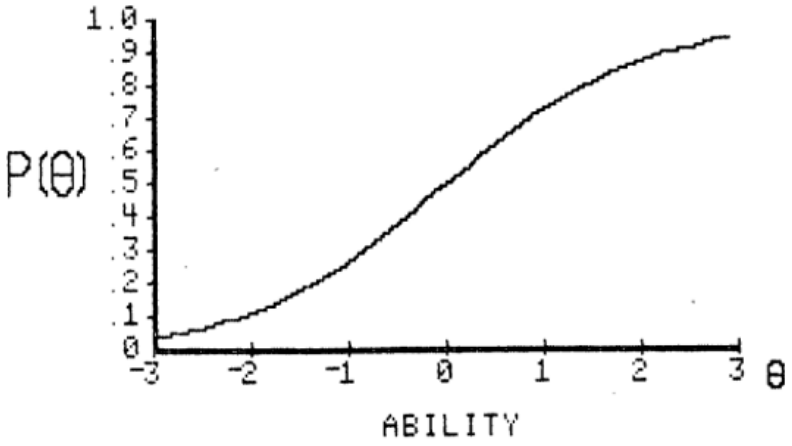

Figure 3.3. A typical item characteristic curve (Baker, 2001, p.7).

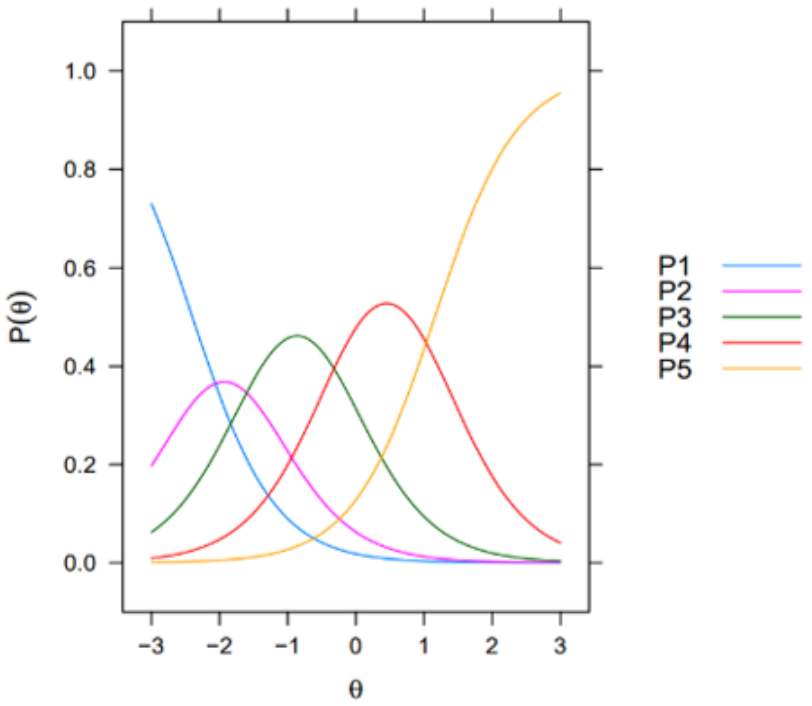

Figure 3.4. ICC for polytomous items.

3.3.3.2. Test Information Function.. The test information function which is summation of single item information functions is one of the highly useful features of Item Response Theory. Compared to a single item information function, the test information function has a significantly higher general level. The test information function provides critical data regarding how well the test is doing in estimating ability across the entire ability score range (Baker *et al.*, 2004).

3.3.3.3. Item Parameters.. The difficulty of the item and discrimination are the two item parameters of IRT (Baker *et al.*, 2004). Firstly, the difficulty of an item is defined as the degree of strength the examinee must possess in order to pass the item (DeVellis, 2017). The difficulty parameter is signified by b. Discrimination, the second parameter, refers to how well an item can differentiate between low and high achievers. The discrimination parameter is denoted by a. Discrimination reflects the slope of the ICC. The steeper curve means the item is able to discriminate better. On the other hand, the flatter of the curve means the item can discriminate less. According to (Baker, 2001), discrimination values between 0.01 - 0.34 are labeled as very low, values between 0.35 - 0.64 are low, values between 0.65 - 1.34 are moderate, values between 1.35 - 1.69 are high, and values higher than 1.70 are very high.

For the scales, the response category function is clearly determined based on the quantity of response categories because polytomous items include more than two response categories (Nering *et al.*, 2011). Unlike 2PL dichotomous IRT models, which describe only one item difficulty parameter, the Graded Response Model (GRM) includes threshold parameters and category boundary for the items based on the number of response categories. X-1 threshold parameters will be given in GRM specifically for an item having X response categories. As an example, n item with five response categories would have four threshold parameters. In Figure 3.4, the item characteristic curves for the item with five response categories, and four threshold parameters can be seen.

3.3.3.4. Logistic Models.. For scales, the graded response model (GRM; Samejima, 1969) and the generalized partial credit model (GPCM; Muraki, 1992) are the two polytomous IRT models that are applied in practice most frequently. The use of GRM in questionnaires has recently gained popularity with common rating scale items in which item responses are ordered categorical responses such as Strongly Disagree (0), Disagree (1), Agree (2), and Strongly Agree (3) (Dai *et al.*, 2021). Through the inclusion of ordered and polytomous item answers, GRM expands the dichotomous two-parameter logistic (2PL) IRT model.

3.3.3.5. Assumptions.. IRT entails three assumptions. The first one is unidimensionality which means the set of items in a test measures a single general latent trait. Factor analysis can be used to investigate unidimensionality. The second one is local independence. The responses of the examinees do not have a statistical relationship with one another under the assumption of local independence. The final one is the shapes of the curves. As mentioned earlier, item characteristic curves show the relationship between the probability of a correct response to an item and the ability scale. In the context of education, for instance, probability of getting the item correct decreases when ability level decreases (Baker *et al.*, 2004; Crocker *et al.*, 2008).

Item Analysis in the Current Study IRT framework was used for item analysis in the current study. Firstly, assumptions were checked. Unidimensionality was investigated by parallel analysis scree plot. One of the important points of developing a scale is that it should measure one general trait. Local independence assumption was checked by Yen's Q3. For monotonicity, item characteristic curves were examined. Secondly, model fits between GRM and Generalized Partial Credit Model (GPCM) were compared. The smaller these fit indices are, the better the models fit the data (Crocker *et al.*, 2008; Hambleton *et al.*, 1991). Thirdly, test information functions were provided by RStudio. Finally, item parameter estimations (item difficulties and item discriminations) and item fit indices were provided. As a result, the items were evaluated, and problematic items were listed by item characteristics curves.

3.3.4. Measurement Invariance

Measurement invariance, known as measurement equivalence, describes the situation in which a psychological tool assesses an unobserved construct in the same way across groups or across time. Since measurement invariance is a prerequisite to compare group means, it takes different forms and is essential for psychological and developmental research (Putnick *et al.*, 2016). Researchers check for invariance among genders (Hong *et al.*, 2003), cultural groups (Senese *et al.*, 2012) and ethnic groups (Glanville *et al.*, 2007). Non-invariance of construct between groups or measurements

can result in faulty inferences about the trial efficacy. Briefly stated, measurement invariance is an essential and pervasive characteristic of psychological and developmental science (Putnick *et al.*, 2016).

There are various steps to test measurement invariance. The majority of tests for measurement invariance report configural, metric, and scalar steps; fewer tests record a residual step (Putnick *et al.*, 2016). According to (Widaman *et al.*, 1997), all these four steps are fundamental to test measurement invariance. Explanations for each step are provided respectively.

Firstly, invariance at the configural level aims to test whether the fundamental organization of constructs is maintained across groups. There are two possibilities after determining configural non-invariance as redefining the construct or assuming that the construct is not invariant, and stopping using group difference and invariance testing. Tests for metric invariance is the next step if configural invariance is supported (Putnick *et al.*, 2016).

Metric invariance is related to the equivalence of item loadings on the factors which means that each item makes a similar amount of contribution to the latent constructs across groups. Metric invariance is examined by constraining factor loadings to be equivalent across groups. The configural invariance model is then compared to the model with constrained factor loadings to assess fit. If the overall model fit is noticeably better in configural invariance model compared to metric invariance model, it indicates that at least one loading is not equivalent across the groups. If the overall model fit is noticeably similar in metric invariance model compared to configural invariance model, it suggests that constraining the loadings across groups has no impact on the model fit (Putnick *et al.*, 2016).

Testing for scalar invariance is the third step if full or partial metric invariance is supported. Scalar invariance is associated with an equivalence of item intercepts. Scalar invariance describes how all mean differences in the shared variance of items are

captured by the latent construct's mean differences. Scalar invariance is checked by constraining the item intercepts to be equivalent across groups. The metric invariance model and the model with constrained item intercepts are compared in order to assess fit. If the overall model fit is noticeably similar in scalar invariance model compared to metric invariance model, it suggests that constraining the item intercepts across groups has no significant effect on the model fit and scalar invariance is supported. Researchers can compare group means on the latent factors after passing the configural, metric, and scalar invariance procedures (Putnick *et al.*, 2016).

Criteria used to evaluate measurement invariance was $\Delta X2$ (Byrne *et al.*, 1989; Marsh *et al.*, 1985; Reise *et al.*, 1993). However, $X2$ is unduly sensitive to small and trivial deviations from a "perfect" model in large samples. For this reason, researchers focus on alternative fit indices rather than $X2$ (Chen, 2007; Cheung *et al.*, 2002; French *et al.*, 2006; Meade *et al.*, 2008). Although (Cheung *et al.*, 2002) criterion of a 0.010 change in CFI for nested models is frequently used, other researchers have proposed several CFI criteria (Meade *et al.*, 2008; Rutkowski *et al.*, 2014) or the use of other fit indices such as ΔTLI and ΔRMSEA (Chen, 2007; Meade *et al.*, 2008; Chen, 2007) proposed a criterion of a 0.010 change in CFI, paired with 0.015 change in RMSEA, for metric invariance or 0.015 for scalar or residual invariance (Meade *et al.*, 2008). Recommended a more cautious cut-off point, which resulted in a 0.002 change in CFI. Model fit was examined in situations comparing 10 or 20 groups by (Rutkowski *et al.*, 2014). It was concluded that 0.020 change in CFI and 0.030 in RMSEA was the most appropriate metric invariance test in large group sizes. Furthermore, 0.010 change in ΔCFI and 0.015 in ΔRMSEA was found the most appropriate one for the tests of scalar invariance. Measurement Invariance in the Current Study It was aimed to test the gender-related measurement invariance in order to understand whether the item characteristics differ between the two response groups and whether differences in item threshold characteristics are reflected or ignored by scale-level analyses. Just for final stage of current study, measurement invariance was examined. Three invariance steps were documented: configural, metric and scalar. Change in the fit values were reported and evaluated.

4. RESULTS

4.1. Results of Pilot Study

In the pilot study, a 15-item scale consisting of 3 factors (perceived usefulness, perceived ease of use and perceived enjoyment) with 5 items in each factor was implemented. Subsequently, collected data was analyzed.

4.1.1. Reliability Results of Pilot Study

Cronbach's Alpha was used to provide evidence for reliability of the test scores in the current study. The scale, overall, exhibited a good level of internal consistency, being exactly at the threshold of 0.80. The Cronbach's Alpha values were calculated for all three subscales and are shown in Table 7. Firstly, it was found that the perceived usefulness factor was below 0.50 limit with a value of 0.46 and displayed unacceptable internal consistency. The perceived ease of use factor showed poor internal consistency with a value of 0.59, falling between the 0.60 to 0.50 range. Finally, the perceived enjoyment factor demonstrated a good degree of internal consistency with a value of 0.89, by exceeding the 0.80 threshold excessively.

In addition to Cronbach's Alpha, McDonald's Omega was used to provide evidence for reliability of the test scores. The difference between values of alpha coefficient and omega were quite small. For overall, McDonald's Omega was found as 0.81 without Item 13.Values of McDonald's Omega were found as 0.45, 0.61 and 0.89 for perceived usefulness, perceived ease of use and perceived enjoyment factors respectively (see Table 4.14).

Table 4.1. The reliability analysis of pilot study.

Factors	**N (Number) (of Items)**	**Cronbach's Alpha**	**Corrected Item-Total Correlation**	**McDonald's Omega**
Perceived Usefulness	5	0.46	(0.12 - 0.34)	0.45
Perceived Ease of Use	5	0.59	(0.16 - 0.53)	0.61
Perceived Enjoyment	5	0.89	(0.65 - 0.79)	0.89
Total	15	0.80	(0.02 - 0.68)	0.81

4.1.2. Validity Results of Pilot Study

4.1.2.1. Content Validity Results.. To provide validity evidence for content validation, table of specification was created and expert opinion was taken. The table of specification which involved dimensions, items in each dimension and numbers provided a fair and representative sample of items and also ensured the alignment between the items and dimensions in the questionnaire. As a result of expert opinion, it was found that 3 items were below critical value of 0.50: Item 2, Item 13 and Item 15 with the values of 0.49, -0.08, and 0.49 respectively. Item 6 was at exactly the threshold. While Item 2 was revised, Item 13 and Item 15 were rewritten. Within all the suggestions of experts about format and indices, according to need, 3 items were revised, 2 items were rewritten and 10 items were decided to remain same. Items before reviewing, index of objective item, change type and items after reviewing are provided in Table 4.2.

Table 4.2. Item Review Statistics and Revisions.

Item No.	**Items Before Reviewing**	**Index of bjective Item**	**Change Type**	**Items After Reviewing**
1.	Ekranda işaretleme yapmak yorucuydu.	0.99	Remain same	Ekranda işaretleme yapmak yorucuydu.

Table 4.2. Item Review Statistics and Revisions. (cont.)

Item No.	**Items Before Reviewing**	**Index of bjective Item**	**Change Type**	**Items After Reviewing**
2.	Biraz önce yanıtladığım testte öğretmenimin yaptığı testlere göre daha az soru çözdüm.	0.49	Revision	Biraz önce yanıtladığım testte, diğer sınavlarıma göre daha az soruyla karşılaştım.
3.	Eğlenceli bir testti.	0.99	Remain same	Eğlenceli bir testti.
4.	Testi çözmeyi sevdim.	0.83	Remain same	Testi çözmeyi sevdim.
5.	Bilgisayar kullanımı ile ilgili öğretmenimden yardım istedim.	0.91	Remain same	Bilgisayar kullanımı ile ilgili öğretmenimden yardım istedim.
6.	Biraz önceki testi çözmek yerine başka bir şey yapmak isterdim.	0.50	Remain same	Biraz önceki testi çözmek yerine, başka bir şey yapmak isterdim.
7.	Biraz önce Yanıtladığım test, öğretmenimin yaptığı testlere göre daha hızlı bitti.	0.75	Revision	Biraz önce yanıtladığım test, diğer sınavlarıma göre daha az zamanımı aldı.

Table 4.2. Item Review Statistics and Revisions. (cont.)

Item No.	**Items Before Reviewing**	**Index of bjective Item**	**Change Type**	**Items After Reviewing**
8.	Soruları ekrandan Okumakta zorlandım.	0.99	Remain same	Soruları ekrandan okumakta zorlandım.
9.	Bilgisayardan okumak, soruları anlamamı zorlaştırdı.	0.99	Remain same	Bilgisayardan okumak, soruları anlamamı zorlaştırdı.
10.	Testin hemen bitmesini istedim.	0.83	Remain same	Testin hemen bitmesini istedim.
11.	Benim bilgi seviyeme uygun sorularla karşılaştım.	0.91	Revision	Benim bilgi Seviyeme daha uygun sorularla karşılaştım.
12.	Soruları bilgisayardan okuyup, işlemleri kağıtta yapmak beni yordu.	0.83	Remain same	Soruları bilgisayardan okuyup, işlemleri kağıtta yapmak beni yordu.
13.	Soruları boş Bırakamadığım için rastgele bir şık işaretlediğim oldu.	-0.08	Rewritten	Soruların boş bırakılamaması beni her soruyu düşünmeye yönlendirdi.
14.	Testi çözmek zevkliydi.	0.99	Remain same	Testi çözmek zevkliydi.

Table 4.2. Item Review Statistics and Revisions. (cont.)

Item No.	**Items Before Reviewing**	**Index of bjective Item**	**Change Type**	**Items After Reviewing**
15.	Biraz önce yanıtladığım testteki sorular, öğretmenimin yaptığı testlerdeki sorulardan daha kolaydı.	0.49	Rewritten	Aşırı zor veya aşırı kolay sorularla karşılaşmadım

4.1.2.2. Construct Validity Results.. In order to provide evidence of construct validity, Confirmatory Factor Analysis (CFA) was conducted.

Whether collected data confirmed factor structure was tested via CFA. The measurement model had 3 factors (perceived usefulness, perceived ease of use and perceived enjoyment) and there were 5 items in each factor.

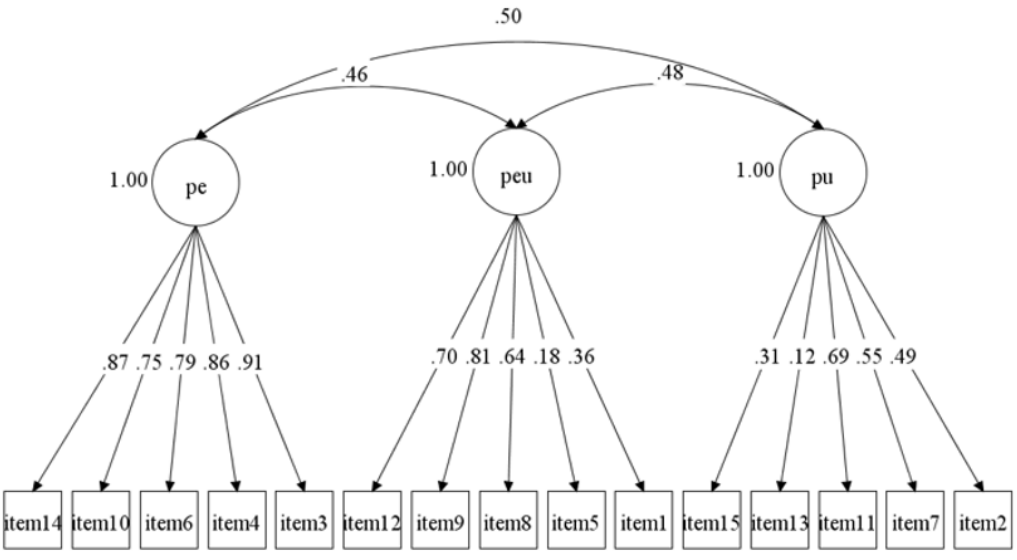

Figure 4.1. CFA diagram for pilot study.

CFI and TLI of pilot study were 0.970 and 0.964 respectively, values over 0.950 indicate a good-fitting model. RMSEA value was calculated as 0.067 which was slightly higher than 0.060. Collected data exhibited good level of fit with measurement model.

The values of standardized factor loadings for Item 1, 5, 13 and 15 were below 0.40 which is the lowest threshold satisfactory fit between subdimensions and respective items (see Table 4.3). Detailed set of results are demonstrated in Table 9.

Table 4.3. The fit ındices in cfa for pilot study.

x^2	**df**	x^2**/dt**	**CFI**	**TLI**	**RMSEA**	**(90% CI)**
3-Factor Model	207.895	87	2.390	0.970	0.964	0.067 (0.055; 0.079)

Table 4.4. The standardized factor loadings for pilot study.

Variables and Items	**Estimate**	**S.E.**	**p-value**
Perceived Usefulness			
Item 2	0.491	0.065	0.000
Item 7	0.548	0.062	0.000
Item 11	0.685	0.074	0.000
Item 13	0.118	0.078	0.133
Item 15	0.308	0.068	0.000
Perceived Ease of Use			
Item 1	0.361	0.068	0.000
Item 5	0.181	0.076	0.018
Item 8	0.643	0.053	0.000
Item 9	0.805	0.047	0.000
Item 12	0.699	0.057	0.000
Perceived Enjoyment			
Item 3	0.913	0.017	0.000
Item 4	0.860	0.018	0.000
Item 6	0.791	0.028	0.000
Item 10	0.746	0.032	0.000
Item 14	0.874	0.020	0.000

4.1.3. IRT Results of Pilot Study

4.1.3.1. Assumptions.. IRT entails three assumptions: unidimensionality, local independence and monotonicity (Baker *et al.*, 2004; Crocker *et al.*, 2008). These assump-

tions were evaluated below. Unidimensionality. The scree plot, represented in Figure 4.2 below and the ratio of first to second eigen values (eigen value #1= 3.81 / eigen value #2= 0.91= 4.19 > 4) indicated that there is one general factor, thus, the unidimensionality assumption was not violated.

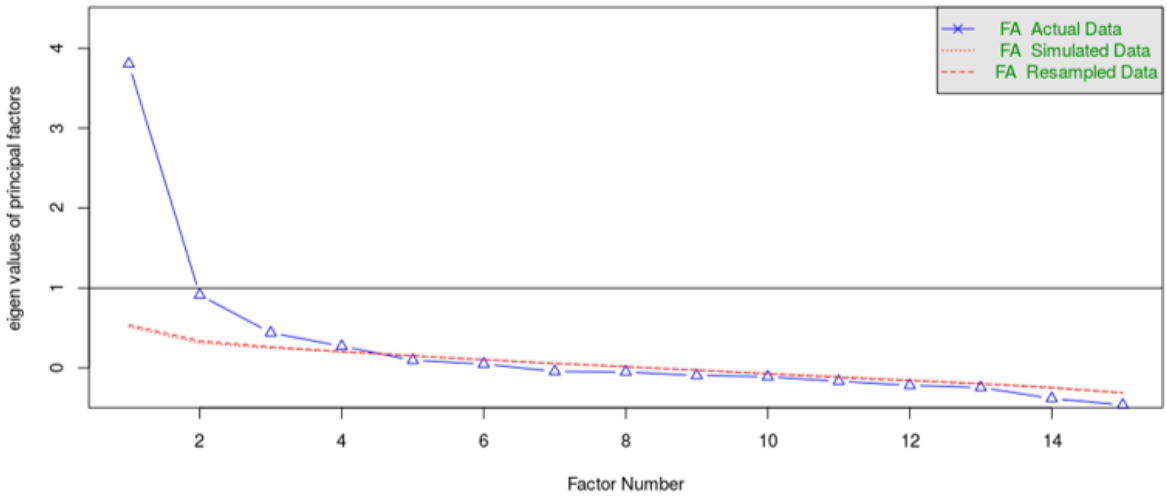

Figure 4.2. Parallel analysis of scree plots for pilot study.

Local Independence to check that the responses of students do not have a statistical relationship with one another under the assumption of local independence, Yen's Q3 Statistic Test was conducted. Local independence is violated by item pairings with values from Yen's Q3 Test greater than 0.300 (Yen, 1984). In current study, item 8 and item 9 exhibited a correlation of 0.512, above threshold of 0.300, and this pairs violated the local independence assumption.

Monotonicity item characteristic curves of 15 item were investigated to check the third assumption and found that monotonicity assumption was not violated. Item characteristic curves for 15 items were provided in Appendix.

4.1.3.2. Test Information Function.. Test information function of pilot study revealed that the initial version of the scale gave more information between the theta values of about -1 and 1.

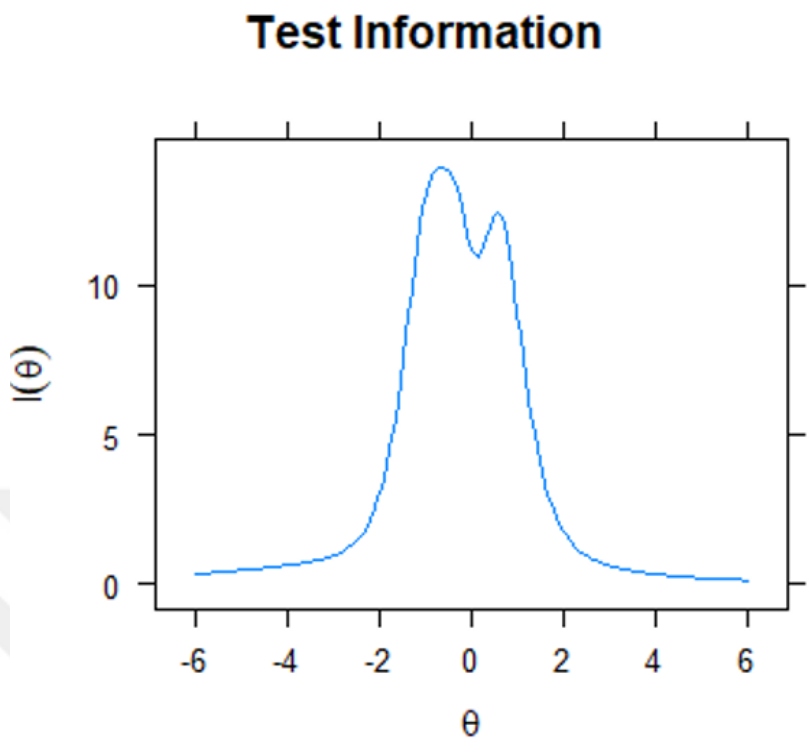

Figure 4.3. Test information function of pilot study.

4.1.3.3. IRT Model Fit.. Graded response model was used to analyze development of a 4-Point Likert Scale for ordered categorical responses from Strongly Disagree (1) to Strongly Agree (4). As a result of comparing model fits, GRM was decided to be used since the smaller fit indices (AIC and BIC values) indicate the data fit the model better.

Table 4.5. Fit indices of pilot study.

Model	**AIC**	**BIC**
Graded Response Model (GRM)	10896.83	11121.22
Generalized Partial Credit Model (GPCM)	10921.60	11145.99

4.1.3.4. Item Parameter.. The item parameter estimation in the current study was presented in Table 4.6. According to Baker's criterion for discrimination values, as an example, item 5 and item 13 were labeled as very low, item 7 and item 8 low, item 9 and item 11 moderate, and finally item 3 and item 4 very high. For the difficulty, item 5 had lowest b value denoted as b1 with a value of -51.135 and item 13 had the highest value with 9.945 in b3.

Table 4.6. The item parameter estimations in pilot study.

Items	**a**	**b1**	**b2**	**b3**
1	0.400	-6.981	-3.041	1.581
2	0.422	-5.827	-3.129	1.483
3	4.051	-0.889	-0.356	0.592
4	3.137	-1.211	-0.525	0.845
5	0.051	-51.135	-28.906	-4.885
6	2.199	-1.047	-0.141	0.804
7	0.455	-5.213	-2.575	0.901
8	0.415	-6.565	-3.516	0.263
9	0.710	-3.302	-1.360	0.730
10	1.949	-0.946	-0.073	1.002
11	0.863	-2.586	-0.938	1.443
12	0.862	-2.145	-0.436	1.183
13	0.081	-27.776	-13.619	9.945
14	3.301	-1.004	-0.381	0.491
15	0.299	-5.825	-1.759	3.884

4.1.4. Overview of Pilot Study Results

According to corrected item-total correlations, CFA factor loadings, IRT item parameters and item characteristic curves; one of the weakest items measuring the defined factors were removed. The final scale with 12 items was determined.

From the Perceived Usefulness factor, Item 13 (Soruların boş bırakılamaması beni her soruyu düşünmeye yönlendirdi.) was removed. Upon examining the values, the lowest corrected item total correlation and lowest factor loading with values of 0.122 and 0.118 respectively belongs to Item 13. In addition to item analysis values (see Table 4.7), ICC of Item 13 confirmed that it should be removed (see Figure 4.4). From the Perceived Ease of Use factor, Item 5 (Bilgisayar kullanımı ile ilgili öğretmenimden yardım istedim.), the most problematic one, was removed. Among all items in this factor, Item 5 had lowest corrected item total correlation and factor loading values with 0.158 and 0.181 in order. In the Perceived Enjoyment factor, ICCs and analysis

values of all 5 items demonstrated no need to remove any items. However, Item 10 (Testin hemen bitmesini istedim.), the weakest item, was removed because of keeping the number of items in factors. As well, the fewer questions the more appropriate for young students.

Table 4.7. Overview of results of pilot study.

Variables and Items	**a**	**b1**	**b2**	**b3**	**Factor Loadings**	**Corrected Item-Total Correlations**
Perceived Usefulness						
Item 2	0.422	-5.827	-3.129	1.483	0.491	0.309
Item 7	0.455	-5.213	-2.575	0.901	0.548	0.335
Item 11	0.863	-2.586	-0.938	1.443	0.685	0.311
Item 13*	0.081	-27.776	-13.619	9.945	0.118	0.122
Item 15	0.299	-5.825	-1.759	3.884	0.308	0.165
Perceived Ease of Use						
Item 1	0.400	-6.981	-3.041	1.581	0.361	0.211
Item 5*	0.051	-51.135	-28.906	-4.885	0.181	0.158
Item 8	0.415	-6.565	-3.516	0.263	0.643	0.483
Item 9	0.710	-3.302	-1.360	0.730	0.805	0.529
Item 12	0.862	-2.145	-0.436	1.183	0.699	0.385
Perceived Enjoyment						
Item 3	4.051	-0.889	-0.356	0.592	0.913	0.786
Item 4	3.137	-1.211	-0.525	0.845	0.860	0.772
Item 6	2.199	-1.047	0.141	0.804	0.791	0.704
Item 10*	1.949	-0.946	-0.073	1.002	0.746	0.653
Item 14	3.301	-1.004	-0.381	0.491	0.874	0.760
* Removed Items						

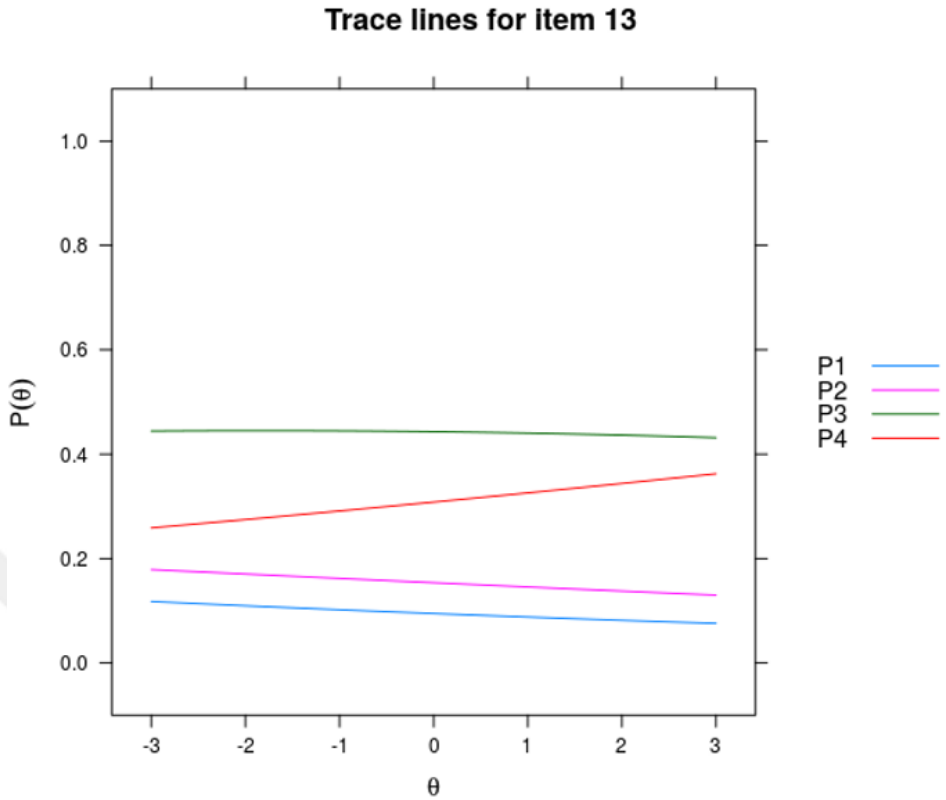

Figure 4.4. ICC for item 13 in pilot study.

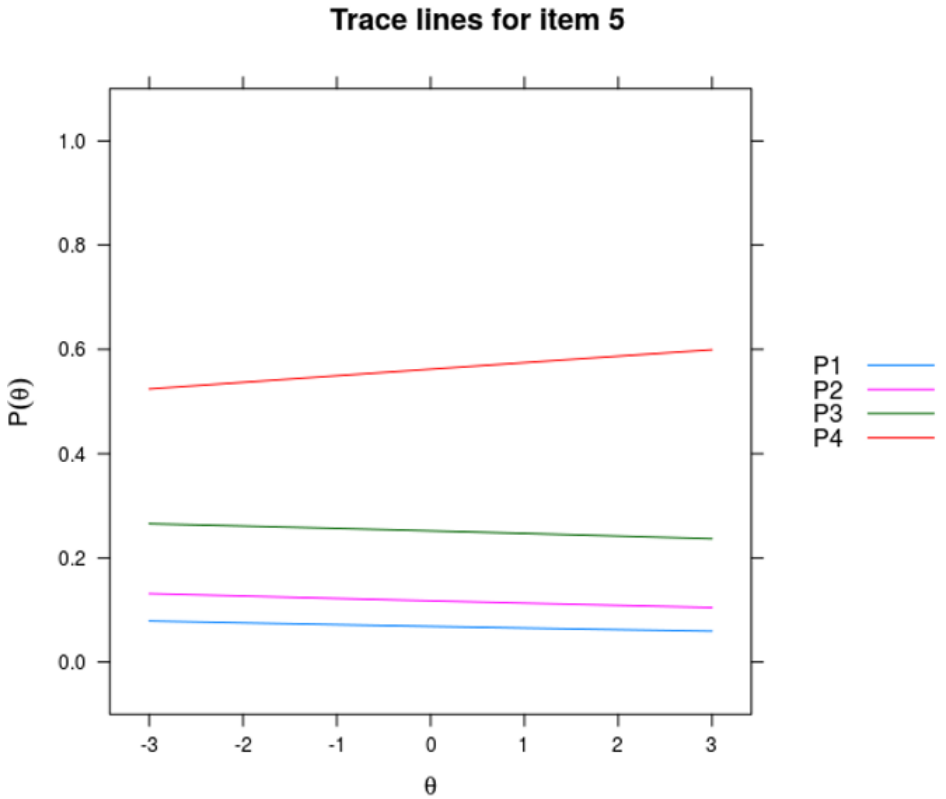

Figure 4.5. ICC for item 5 in pilot study.

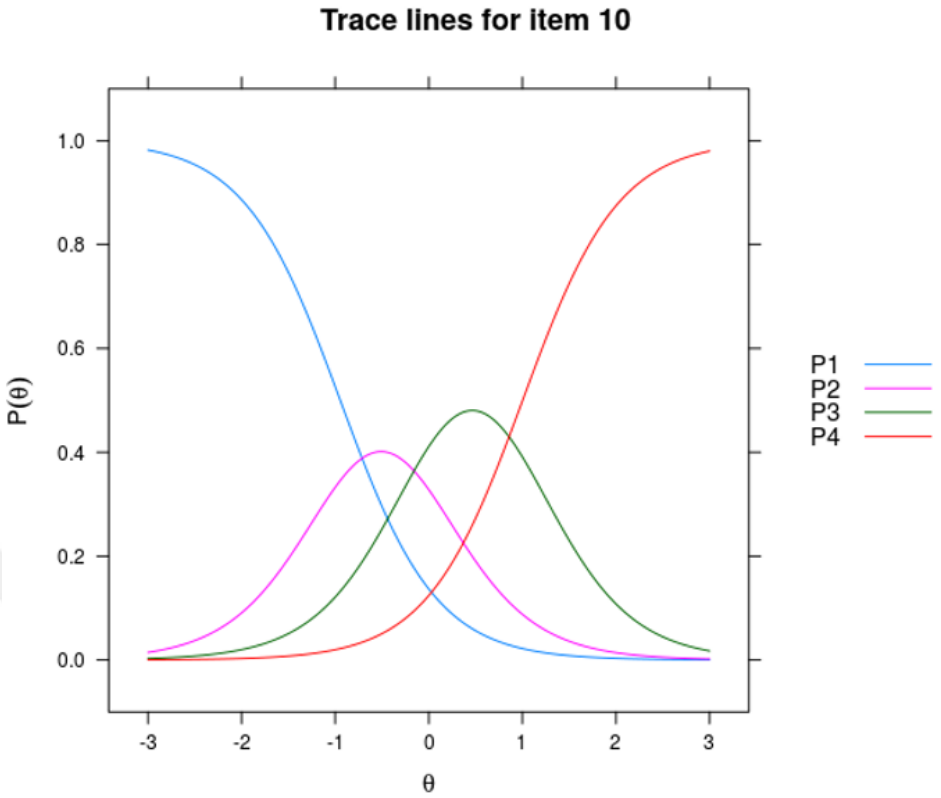

Figure 4.6. ICC for item 10 in pilot study.

4.2. Results of Final Study

In the final study, a 12 item-scale consisting of 3 factors (perceived usefulness, perceived ease of use and perceived enjoyment) with 4 items in each factor was implemented. Data collected from the final study were analyzed following procedures similar to how data analyzed in pilot study.

4.2.1. Reliability Results of Final Study

It was found that the scale had, overall, a good level of internal consistency, exactly at the 0.80 cutoff. For each of the three subscales, the Cronbach's Alpha and McDonald's Omega values were computed. Firstly, the perceived usefulness factor exhibited poor internal consistency for both Cronbach's Alpha and McDonald's Omega value of 0.52. The perceived ease of use factor showed questionable internal consistency, being preciously at the threshold of 0.60 for Cronbach's Alpha and 0.62 for McDonald's

Omega. Lastly, the perceived enjoyment factor demonstrated a good degree of internal consistency with a value of 0.82 for Cronbach's Alpha and 0.83 for McDonald's Omega, by surpassing the 0.80 threshold (see Table 4.8).

Table 4.8. The reliability analysis for final study.

Factors	N (Number of Items)	Cronbach's Alpha	Corrected Item-Total Correlation	McDonald's Omega
Perceived Usefulness	4	0.52	(0.27 - 0.33)	0.52
Perceived Ease of Use	4	0.60	(0.21 - 0.47)	0.62
Perceived Enjoyment	4	0.82	(0.48 – 0.74)	0.83
Total	12	0.80	(0.18 – 0.68)	0.79

4.2.2. Validity Results of Final Study

4.2.2.1. Construct Validity Results.. In order to provide evidence of construct validity, Confirmatory Factor Analysis (CFA) was conducted. CFI and TLI were found as 0.969 and 0.959 in order, values greater than 0.950 indicate a good-fitting model. RMSEA value was calculated as 0.067, exceeding but pretty close to 0.060 cutoff.

Based on the obtained CFI, TLI and RMSEA values, it was determined that the collected data exhibited a good level of compatibility with proposed measurement model (see Table 4.9). The value of standardized factor loadings for Item 1 was 0.27 less than 0.40 which is the lowest threshold satisfactory fit between subdimensions and corresponding items. A comprehensive collection of findings is shown in Table 16.

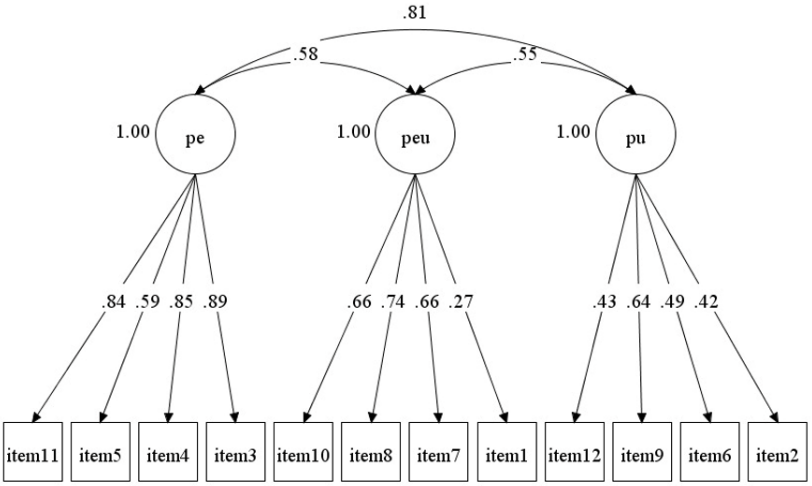

Figure 4.7. CFA diagram for final study.

Table 4.9. The fit indices in cfa for final study.

	x^2	**df**	x^2	**CFI**	**TLI**	**RMSEA (90% CI)**
3-Factor Model	106.666	51	2.091	0.969	0.959	0.067 (0.049; 0.085)

Table 4.10. The fit indexes.

Variables and Items	**Estimate**	**S.E.**	**p-value**
Perceived Usefulness			
Item 2	0.423	0.066	0.000
Item 6	0.491	0.070	0.000
Item 9	0.635	0.060	0.000
Item 12	0.429	0.064	0.000
Perceived Ease of Use			
Item 1	0.274	0.076	0.000
Item7	0.657	0.055	0.000
Item 8	0.737	0.062	0.000
Item 10	0.656	0.064	0.000
Perceived Enjoyment			
Item 3	0.891	0.021	0.000
Item 4	0.848	0.024	0.000
Item 5	0.588	0.051	0.000
Item 11	0.843	0.028	0.000

4.2.3. IRT Results of Final Study

4.2.3.1. Assumptions.. Unidimensionality. The parallel analysis scree plot, demonstrated in Figure 4.8 and the ratio of first to second eigen values (eigen value #1= 3.32 / eigen value #2= 0.65= 5.11 > 4) proved that there is one general factor, hence, unidimensionality assumption was not violated.

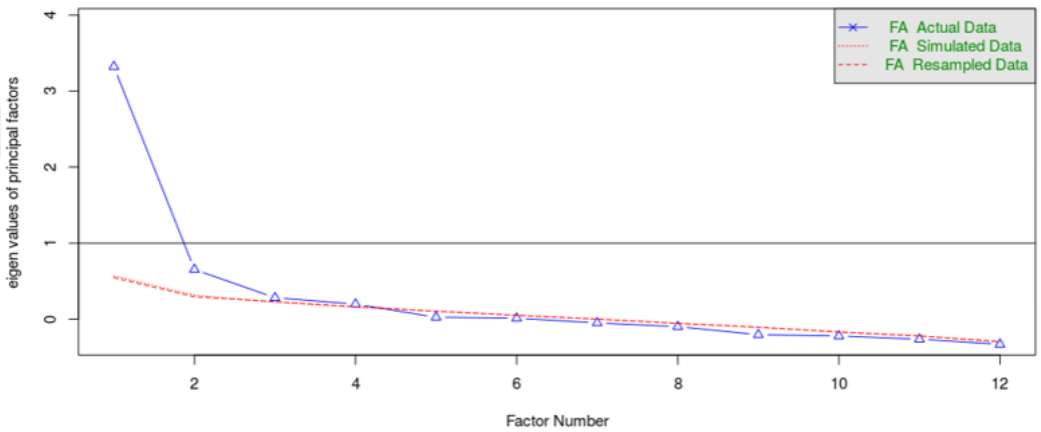

Figure 4.8. Parallel analysis scree plots for final study.

Local Independence yen's Q3 Statistics Test was conducted to test local independence assumption. Since there were item pairs which exhibited item correlation above 0.300, see Appendix, local independence assumption was violated.

Monotonicity item characteristic curves of 12 items were examined and found that third assumption of IRT was not violated. Item characteristic curves for 12 items were provided in Appendix.

4.2.3.2. Test Information Function.. Test information function of final study revealed that the final version of the scale gave more information between the theta values of about -1 and 1.

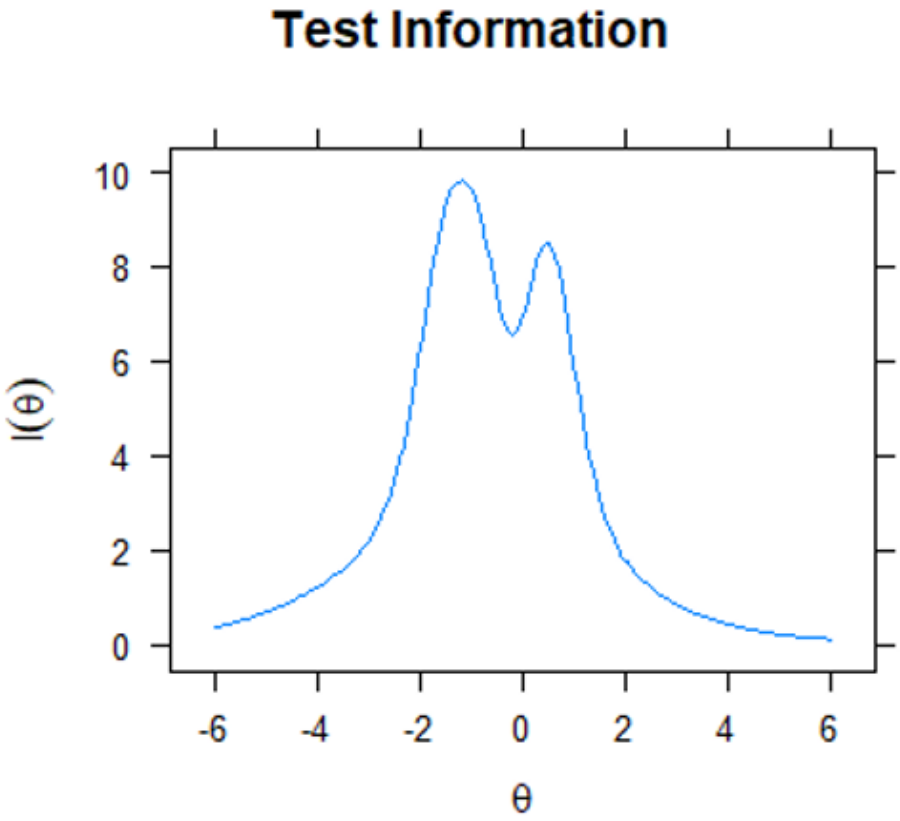

Figure 4.9. Test information function of final study.

4.2.3.3. IRT Model Fit.. When the models are compared, it was deemed appropriate to use GRM because it had lower AIC and BIC values.

Table 4.11. Fit indices for final study.

Model	**AIC**	**BIC**
Graded Response Model (GRM)	6477.09	6498.39
Generalized Partial Credit Model (GPCM)	6644.36	6665.66

4.2.3.4. Item Parameter.. The item parameter estimation in the final study was presented in Table 4.12. In line with Baker's assessment, item 1 had low discrimination which was the lowest among all items. Items with different degrees of discrimination were included in current scale. For the difficulty, item 1 had lowest b value denoted as b1 with a value of -8.093 and item 1 also had the highest value with 2.690 in b3.

Table 4.12. The fit indexes.

Item	a	b1	b2	b3
1	0.287	-8.093	-4.838	2.690
2	0.725	-2.797	-0.957	2.305
3	3.092	-1.436	-0.849	0.503
4	2.672	-1.752	-1.009	0.671
5	1.279	-2.146	-1.103	0.650
6	0.797	-2.872	-1.436	1.298
7	0.780	-3.491	-1.989	0.717
8	0.931	-2.607	-1.437	0.619
9	1.216	-2.353	-0.922	1.328
10	0.908	-2.799	-1.327	0.865
11	3.081	-1.424	-0.929	0.402
12	0.781	-3.742	-1.403	1.188

4.2.4. Measurement Invariance Results of Final Study

The goodness of fit statistics for gender-related measurement invariance was presented in Table 4.13. Configural invariance model demonstrated a good model fit, X^2 (102)= 193.521, CFI = 0.950, RMSEA= 0.086, except for RMSEA since exceeding 0.060 threshold. However, since good model fit was found in other model fit indices, it was concluded that the results provide sufficient support for equal number of factors between males and females. After evidence for configural invariance across gender was found, the viability of the metric invariance model was tested. Then, the results showed strong support for metric invariance, with item factor loading equivalence constraints producing change in model fit indices close to cutoff values, ΔCFI= 0.007, ΔRMSEA= 0.009. Finally, scalar invariance was examined and results demonstrated support for scalar invariance with item intercepts constraints, ΔCFI= 0.001 ΔRMSEA= 0.007. Overall, it was concluded that examinees' attitudes toward CAT have similar meaning across gender groups.

Table 4.13. The fit indexes.

Model	X2	df	X2/df	CFI	RMSEA	CFI	RMSEA
Configural	193.521	102	1.897	0.950	0.086	-	-
Metric	190.183	111	1.713	0.957	0.077	0.007	0.009
Scalar	210.066	132	1.591	0.958	0.070	0.001	0.007

4.2.5. Descriptive Statistics of Final Study

The descriptive statistics was reported to understand how students perceived CAT. Descriptive Statistics for the Perceived Usefulness Factor. This part of the analysis presents the findings regarding the degree which an examinee believes that CAT is useful compared to traditional paper-pencil test. This investigation was responded with the scores obtained from the perceived usefulness factor. The mean score for usefulness of CAT perceived by examinees was 11.42 and the standard deviation was 2.31. The lowest and highest scores were 4.00 (n= 1) and 16.00 (n= 7). Figure 4.10 represents the distribution of scores. The mean of perceived usefulness was close to 12 which showed that students agreed that CAT was useful.

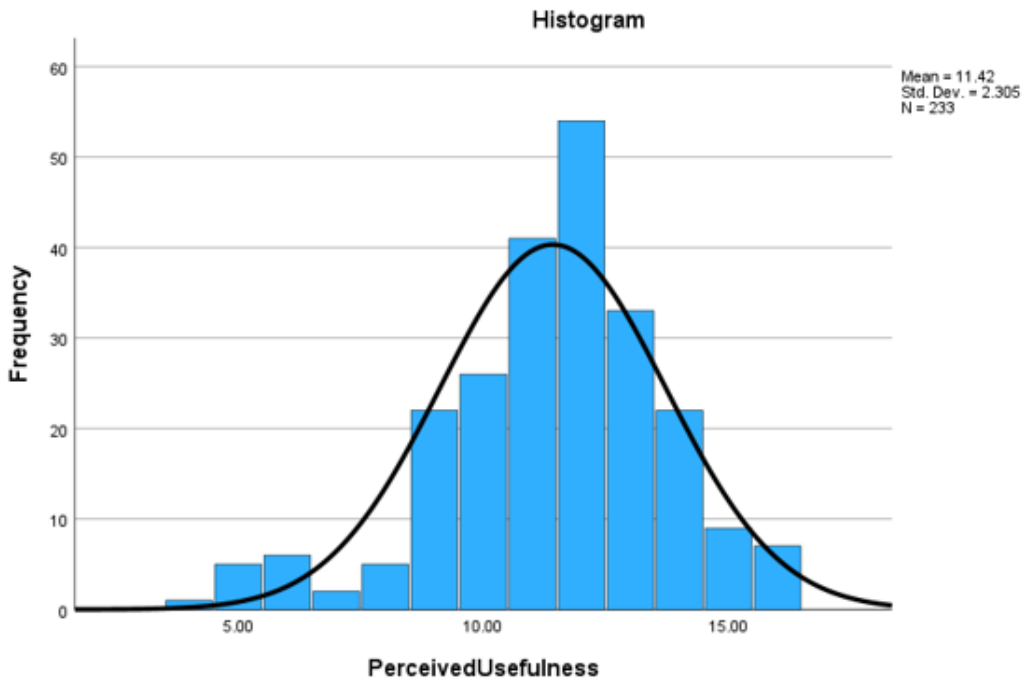

Figure 4.10. Histogram of perceived usefulness scores.

Descriptive Statistics for the Perceived Ease of Use Factor. This part of analysis demonstrates the findings regarding the degree of an examinee beliefs of using CAT being free of any effort. The mean score for ease of use perceived by examinees was 12.11. The minimum score for the perceived ease of use factor was 4.00 (n= 1) and the maximum score was 16.00 (n= 15). The distribution of the perceived ease of use scores was demonstrated in Figure 4.11. The mean of perceived ease of use was larger than 12 which showed that students agreed that CAT was easy to use.

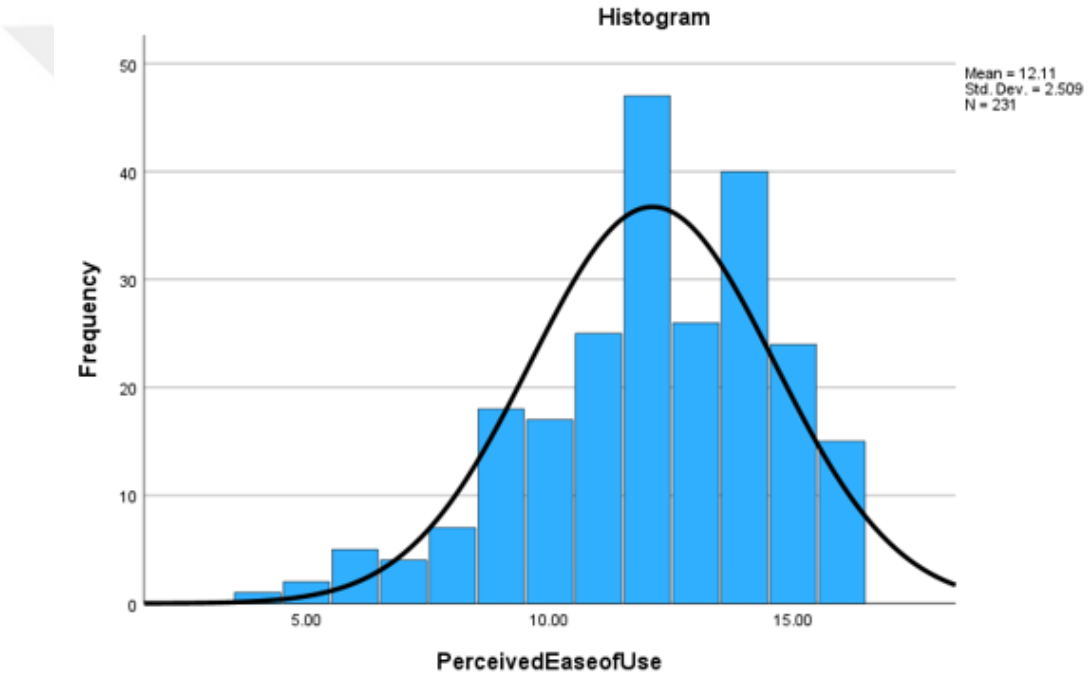

Figure 4.11. Histogram of perceived ease of use scores.

Descriptive Statistics for the Perceived Enjoyment Factor. Within this part, results regarding participants' overall satisfaction with the CAT system was presented in the scope of the perceived enjoyment factor. The mean score for enjoyment perceived by examinees was 12.08. While the minimum score for the perceived enjoyment factor was 4.00 (n= 4), the maximum score was 16.00 (n= 28). Figure 4.12 shows the distribution of the perceived ease of use scores. Similarly, students agreed that CAT was enjoyable.

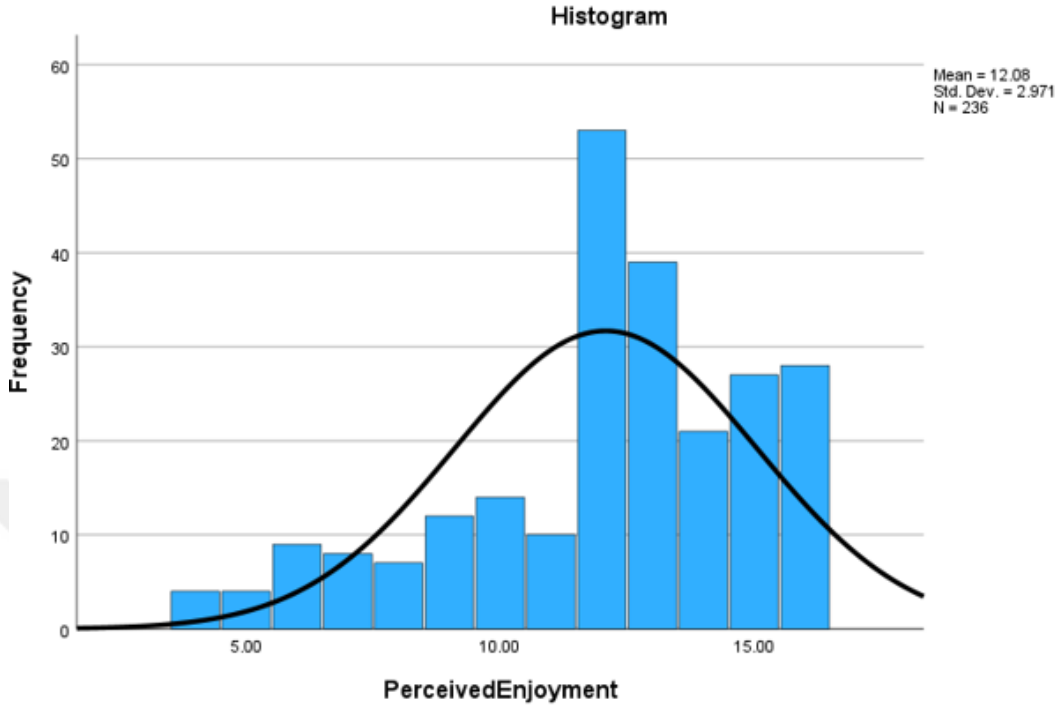

Figure 4.12. Histogram of perceived enjoyment scores.

Table 4.14. Descriptive statistics.

Factor	**N (Valid)**	**Min.**	**Max.**	**Mean**	**Std. Dev.**
Perceived Usefulness	233	4.00	16.00	11.42	2.31
Perceived Ease of Use	231	4.00	16.00	12.11	2.51
Perceived Enjoyment	236	4.00	16.00	12.08	2.97

5. CONCLUSION

The core objective of the current study was to develop a scale examining test-takers' attitudes toward CAT. Within this motivation, a three-factor measurement model was proposed. TAM provided the conceptual framework for the research model. The factors of perceived usefulness and perceived ease of use were derived from TAM, while the perceived enjoyment had its foundation in literature. The study was carried out in two phases: pilot and final study. Throughout both phases, following a live CAT in a 4th grade mathematics test, students responded to the scale. In the pilot phase, a 15-item initial scale was administered and gathered data was analyzed. Eventually, one item was removed from each factor, resulting in a final scale of 12 items. The final form was implemented once more for gathering evidence for reliability and validity of the scores. There is a limited number of scales measuring examinees' attitudes toward CAT in the literature, notably within the field of education. The significance of this scale lies in providing the literature with a tool measuring examinees' attitudes toward CAT. The scale developed in current study offers a more profound comprehension of the test-takers' attitudes toward CAT, then ultimately, future CAT implementations might be truly advanced.

5.1. Significance of the Scale

When looking at the history of CAT, it's seen to date back to the 1960s. ASVAB (Armed Services Vocational Aptitude Battery) is known as one of the first developed adaptive tests, enabling selection and categorization within the armed services (Lord, 1952; Rasch, 1960; Wright *et al.*, 1979). Later on, CAT evolved into an assessment system immensely used in diverse procedures such as military environments, health processes, recruitment processes, and employee evaluation. Studies conducted by English, (Reckase *et al.*, 1977; Bejar *et al.*, 1978) indicate that CAT started to have repercussions in educational testing. Subsequently, CAT applications continued to proliferate in the field of education without slowing down. In this point, investigating examinees'

attitudes toward CAT has become the most remarkable need to refine and optimize adaptive testing in the educational context. However, scales measuring test-takers' attitudes toward CAT are limited. In the meantime, when looking closer, existing scales mostly administered after CAT experiences in fields other than education such as health sector (eg. Nikolaus *et al.*, 2014; Chuesathuchon, 2008), one of the most comprehensive CAT-related studies in the educational field, extensively investigated CAT process in his thesis titled "Computerized Adaptive Testing in Mathematics for Primary Schools in Thailand". The researcher has designed the infrastructure of the entire CAT experience both digitally and theoretically, also, developed scale measuring students' attitudes toward CAT. The scale was conceptualized from five aspects and consisted of totally 30 items as follows: Like and Interest ("I like the computerized adaptive test because of its immediate feedback"), Confidence with and Use of CAT (I took the computerized adaptive test with confidence), CAT as Modern and Useful ("Computerized adaptive testing saves money"), CAT is Reliable ("CAT gives reliable results") and CAT Recommendations ("I will tell my friends about CAT"). Regarding the findings of study, there was acceptable overall fit. Similar aspects between scale developed in current study and the scale developed by (Chuesathuchon, 2008) are in the field of education, even in mathematics, and based on live CAT experience. While both scales bear resemblance they also differ in a few but pretty important points. The most outstanding distinction lies in informing students about that they experience live-CAT during the administration. Clearly saying, in the study of (Chuesathuchon, 2008), there were items outspokenly referring to CAT characteristics. Students knew that they took live CAT mathematics test and filled out the questionnaire with this awareness. Nevertheless, in the current study, students were not aware of the fact that they attended live CAT, hence, the concept of "CAT" was not mentioned in scale items. Students respond the questionnaire based solely on perceived experience without prior technical or practical knowledge about CAT.

Regarding literature, notwithstanding the rarity of scale development studies, examinees' attitudes toward CAT was investigated in certain studies (Daphine *et al.*, 2020; English, Reckase *et al.*, 1977; Legg *et al.*, 1992; Moe *et al.*, 1988; Özbaşı, 2016;

Schmidt, Urry *et al.*, 1978; Tonidandel *et al.*, 2000; Vispoel, Rocklin *et al.*, 1994). Within most of the studies, the thoughts of students regarding CAT have been collected haphazardly with different formats. Instruments for data collection was developed in the light of expert opinion and literature (Cheusathuchon, 2008). Suggested five aspects based on the literature. Schmidt, (Urry *et al.*, 1978), for instance, wrote fill in the blank questions asking what the worst thing in CAT, then, responses were evaluated based on frequencies. Examinees had an opportunity to express their thoughts in details and, essentially, the findings of the study provided a broad approach regarding the opinions of examinees toward CAT. In the study of (Özbaşı, 2016), the instrument was developed based on expert opinion and literature. The context and structure of questions and also results from previous research considerably inspired the current study, principally, while searching the characteristics of CAT examinees emphasized most. Once for all, in current study, the scale was developed by taking TAM as the theoretical framework of the study and by extensively scanning the literature and expert opinion.

5.2. Dimensions of the Scale

By integrating a motivational approach into TAM, theoretical framework of current study captured both extrinsic (perceived usefulness and perceived ease of use) and intrinsic (perceived enjoyment) motivators for measuring examinees' attitudes toward CAT. Over the last 25 years, the TAM widely utilized as theoretically justified model to comprehend how users accept or reject any technology, in addition to determine factors influencing the intention to use any technology (Alharbi *et al.*, 2014; Chen *et al.*, 2011; Lee, Cheung *et al.*, 2005; Eraslan Yalçın *et al.*, 2019; Teo *et al.*, 2011). Analysis of TAM critically and comparisons with other intention-based models, including Theory of Reasoned Action (TRA) and Theory of Planned Behavior (TPB), have concluded that TAM is a theoretically tailored for computer-technology acceptance thanks to great research importance in IT (Lee *et al.*, 2005). In the field of educational research, TAM has been frequently adopted to e-learning platforms such as Moodle (eg. Eraslan Yalçın *et al.*, 2019), CBA (Alkış, 2010; Terzis *et al.*, 2011a) and web-based assessment (Dizon, 2016). There is rare CAT-related research, solely in health sector, based on

TAM (Nikolaus *et al.*, 2014). In the study of (Nikolaus, 2014) aimed to test the usability of CAT for a kind of disease, and findings revealed that TAM was important directory to test the usability of the CAT. In fact, in current study, TAM was extended by the perceived enjoyment. Regarding the literature, TAM has been extended, named as E-TAM, by integrating different external variables such as social norm, user interface design and computer anxiety, computer self-efficacy (Eraslan Yalçın *et al.*, 2019). In the field of education, there are also several studies integrating perceived enjoyment in TAM (Karagöz, 2022; Lee *et al.*, 2005; Martínez-Torres *et al.*, 2008; Mun *et al.*, 2003; Saadé *et al.*, 2005). Similarly, in current study, perceived enjoyment was included, although there are other variables that can be directly associated with CAT such as perceived fairness, computer anxiety and self-efficacy. Perceived fairness, especially, might be posit a significant impact on students' attitude toward CAT, because there is a strong procedural justice literature (eg. Tonidandel and Quinones, 2000) that proves violations of justice rules such as differences in both number of questions and test duration results in negative reactions toward CAT. In current study, inclusion of perceived enjoyment factor was rooted from suggestions of the (Carroll *et al.*, 1988; Diagnostic Research, 1988; Webster, 1989) and Davis *et al.* 1992; Carroll *et al.*; 1988), similar to Diagnostic (Research, 1988), advocated that enjoyment is an essential factor underlying user acceptance of any technology. Then, (Webster, 1989) strongly proved that computer use encourages playfulness. However, whether enjoyment has an effect on behavioral intentions had not been investigated until the study of (Davis *et al.*, 1992). The study of (Davis *et al.*, 1992) investigated perceived usefulness and perceived enjoyment, then found that both influenced entirely behavior intention. Nonetheless, usefulness was almost four to five times more influential than enjoyment in determining user behavioral intentions.

5.3. Experience of Examinees

The results of current study were consistent with prior studies that examinees' attitudes toward CAT were overwhelmingly positive (Baghi *et al.*, 1991; Legg *et al.*, 1992; Moe *et al.*, 1988; Schmidt *et al.*, 1978; Vicinio *et al.*, 1997). Students found

CAT useful, easy to use and enjoyable. It should not be overlooked that number of factors such as age, test field, pre-information about CAT and physical conditions of test environment have greatly impacted CAT experience of attitudes.

Chuesathuchon (2008) found that primary school students had a very positive attitude towards CAT for mathematics. Even though participants were young, comprehensive evaluation were expected with scale items such as "CAT saves money" and "CAT gives reliable results". Students agreed most that CAT is very useful, on the other hand, did not endorse most that they took CAT with confidence (Schmidt *et al.*, 1978). Conducted content analysis of responses for open-ended questions and revealed several critical findings. In the study of (Schmidt *et al.*, 1978), considering the CAT experience, "required less time" was most given response (35%) to the question of what the best characteristics of CAT, whereas "fewer items are administered" was the last one (1%) for the same question. In fact, 2 percent of participants respond the question as "items match to examinee ability". Within the scope of perceived usefulness factor in current study, time efficiency, fewer questions and not too easy or too difficult items in CAT comparing to P and P was measured. The most common answer (23%) to the question asking about the worst feature of CAT was "examinee can not go back and review answer". The finding strongly supported the literature, inability to skip or review answers was criticized by participants frequently and researchers/practitioners concluded that this characteristic led to negative attitudes toward CAT. Within the scope of perceived usefulness factor, in current study, mentioned on-target points were referred in items. According to results, overall students believed that CAT was needed to shorter amount of time, fewer questions and did not present too easy or too difficult questions (Schmidth *et al.*, 1978). Also addressed ease of CAT use, 6% and 13% of participants respectively listed "negative attitudes towards computers general" and "problems in reading the screen" as the worst thing of CAT. Consistently, in the study of (Özbaşı, 2016), 32.8% of students expressed that marking on screen was exhausted. In fact, 65.2% of students agreed that items can be readable on screen easily while 88.2% of them endorsed that items can be readable in P and P test easily. This finding was clarified with daily computer use, means that, students who spent time on com-

puter more than 2 hours in a day did not have difficulties in reading questions on screen or marking. This exploration was supported with another research (Boyle, 1997; Lilley, 2007; Madsen, 1986; Pino-Silva, 2008; Segall and Moreno, 1999). However, in current study, student agreed that CAT was easy to use. The result fails to align with earlier research might be explained by increasingly widespread use of computers, tablets and phones among young students, correspondingly, increased familiarity with using those technological tools. Finally, test-takers enjoyed doing CAT (Chuesathuchon, 2008). The findings of current study supported the literature that students agreed that CAT was enjoyable.

5.4. Mathematics and Science Education and CAT

CAT implementations have increased worldwide and suggested as an alternative instrument to P and P test (Desa *et al.*, 2007). CAT has some advantages, psychologically proved that CAT lessens the anxiety of the examinees because each test taker is challenged at their own level (Cheusathuchon, 2008; Eggen *et al.*, 2006; Van der Linden *et al.*, 2009). Low achievers do not have to respond tests which are too difficult for their ability level or too long. There are also some disadvantages of CAT such as inability to skip or review items. Nonetheless, overall, the attitudes of examinees toward CAT were overwhelmingly positive (Baghi *et al.*, 1991; Legg *et al.*, 1992; Moe *et al.*, 1988; Schmidt *et al.*, 1978; Vicinio *et al.*, 1997).

In parallel to other fields, in mathematics and science, CAT administrations have been used, ranging from large scale international tests such as PISA to in-class assessments (Yamamoto *et al.*, 2019; Chinn *et al.*, 2001) argue that assessment procedure is fundamental to measure scientific and mathematical reasoning. In addition, students frequently have a concern about grading and mathematics anxiety, then develop negative attitudes toward mathematics (Yenilmez, 2007). In other words, mathematics achievement might be related to mathematics anxiety (Şad *et al.*, 2016) and attitudes toward mathematics (Yenilmez, 2007). At this point, to break this chain, new administrations in measurement such as CAT might be a part of solutions. Within science, it

was advocated starting taking advantage of CAT by integrating in teaching in science classes (Papanastasiou, 2003). Furthermore, in the study of (Samsudin *et al.*, 2020), item bank of TIMSS was calibrated for CAT, and concluded that students showed positive acceptance in technology use for assessment. The key finding was that the CAT implementation presented a lot of data about such as curriculum and instructional practices useful for policymakers, researchers, curriculum developers and educators.

Focusing on Turkey within mathematics achievement, in national and international assessment, mathematics is one of the most unsuccessful courses (Afacan *et al.*, 2023). For instance, considering national exams such as TYT and AYT, according to statistics provided by Student Selection and Placement Center in 2022, the mathematics course has the lowest correct average following the science course. Similarly, in international exams, in PISA 2022, according to the PISA 2022 Turkey report provided by Ministry of Education (MEB), while the average score of the participating OECD countries in mathematics was 472, Turkey remained below the average with 453 points. What has become clear is that mathematics achievement is moderately low in Turkey. At this point, experiencing CAT might provide both solution and extensive information. Profits of CAT such as shorter amount of testing time and a higher level of precision in test scores might be opportunities for mathematics and science courses to have more time for instruction, detect students' abilities accurately, thus pave the way for personalized learning.

Spreading CAT is important. However, CAT has both advantages and disadvantages. Characteristics of CAT has an impact on examinees' attitudes and behavioral intention toward CAT and to discover examinees' attitudes are crucial to enhance future CAT implementations more and make sensible decisions about implementations in the long-term. Herein, the scale developed might be used and provided benefit.

5.5. Limitations

There is a limitation of the current study. In the pilot of the study, data was exclusively collected from private schools since public schools in primary level have not computer laboratories. Nonetheless, the vast majority of the students have been enrolled in public schools in Turkey. According to 2021-2022 data presented by Ministry of Education (MEB), while 8 percent of students attended in private schools, 82 percent of students attended public schools. Considering this ratio, collecting data just from students in private schools might not completely reflect the state of affairs in Turkey.

5.6. Suggestions for Further Research

The first suggestion, for further studies, is to apply the scale following CAT experiences in different courses and also different grade levels. In current study, the item pool was prepared based on 4th-grade mathematics objectives. Hence, the scale administration was limited to 4th grade and 5th grade at the level of readiness in mathematics achievement tests.

The second suggestion is examining the relationship between examinees' attitudes toward CAT and test scores on CAT. Additionally, the relationship between external variables such as perceived fairness, computer anxiety and self-efficacy frequently used to measure end-users' attitudes toward any technology might be examined.

REFERENCES

Afacan, P. and M. A. Bircan, 2023, "İlkokul Öğrencilerinin Matematik Dersindeki Başarısızlık Nedenlerinin Sınıf Öğretmenlerinin Görüşlerine Göre Değerlendirilmesi", *Amasya Üniversitesi Eğitim Fakültesi Dergisi*, Vol. 12, No. 1, pp. 1-17.

Ajzen, I and M. Fishbein, 1980, *Understanding Attitudes and Predicting Social Behavior*, Englewood Cliffs, Prentice-Hall, New Jersey.

Alharbi, S. and S. Drew, 2014, "Using the Technology Acceptance Model in Understanding Academics' Behavioral Intention to Use Learning Management Systems", *International Journal of Advanced Computer Science and Applications*, Vol. 5, No. 1, pp. 143-155.

Alkış, N., 2010, *Identifying Factors that Affect Students' Acceptance of Web-Based Assessment Tools Within the Context of Higher Education*, M.S. Thesis, Middle East Technical University.

American Educational Research Association, American Psychological Association (APA), and National Council on Measurement in Education, 1999, *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, District of Columbia.

Baghi, H., R. Gabrys and S. Ferrara, 1991, *Applications of Computer-Adaptive Testing in Maryland*, Paper presented at the annual meeting of the American Educational Research Association, Street, Washington, District of Columbia.

Baghi, H., S. Ferrara and R. Gabrys, 1992, *Student Attitudes Toward Computer Adaptive Test Administrations*, Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Lousiana.

Baker, F., 2001, *The Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, Washington, District of Columbia.

Baker, F.B. and S. H. Kim, 2004, *Item Response Theory: Parameter Estimation Techniques*, Chain Reaction Cycles, London.

Bejar, I. I., and D.J. Weiss, 1978, "A Construct Validation of Adaptive Achievement Testing", *Psychometric Methods Program Research Report*, Vol. 1, No: 1, pp. 78-94.

Bennett, R.E., 2002, "Inexorable and Inevitable: The Continuing Story of Technology and Assessment", *Journal of Technology, Learning and Assessment*, Vol. 1, No. 1, pp. 1-24.

Birnbaum, A., 1968, "Some Latent Trait Models", *Statistical Theories of Mental Test Scores*, Vol. 2, No:1, pp. 397-479.

Blais, J.G. and G. Raiche, 2002, *Some features of the Sampling Distribution of the Ability Estimate in Computerized Adaptive Testing According to Two Stopping Rules*, Paper presented at International Objective Measurement Workshop, New Orleans, Lousiana.

Browne, M.W. and R. Cudeck, 1993, "Alternative Ways of Assessing Model", *Testing Structural Equation Models*, Vol. 4, No: 1, pp. 137-162.

Boyle, T., 1997, "Design for Multimedia Learning", *Computing Reviews*, Vol. 38, No. 9, pp. 371-389.

Burke, M.J., J. Normand and N.S. Raju, 1987, "Examinee Attitudes Toward Computer-Administered Ability Testing", *Computers in Human Behavior*, Vol. 3, No. 2, pp. 95-107.

Byrne, B.M., R.J. Shavelson and B. Muthén, 1989, "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance", *Psychological Bulletin*, Vol. 105, No. 3, pp. 456-466.

Campbell, D.T. and D.W. Fiske, 1959, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix", *Psychological Bulletin*, Vol. 56, No. 2, pp. 81-105.

Carmines, E.G. and R.A. Zeller, 1979, *Reliability and Validity in Assessment*, Thousand Oaks, California.

Carroll, J.M. and J.C. Thomas, 1988, "Fun. ACM SIGCHI Bulletin", *ACM New York*, Vol. 19, No. 3, pp. 21-24.

Chen, F.F., 2007, "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance", *Structural Equation Modeling*, Vol. 14, No. 3, pp. 464-504.

Chen, S.C., L. Shing-Han and L. Chien-Yi, 2011, "Recent Related Research in Technology Acceptance Model: A Literature Review", *Australian Journal of Business and Management Research*, Vol. 1, No. 9, pp. 124.

Cheung, G.W. and R.B. Rensvold, 2002, "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance", *Structural Equation Modeling*, Vol. 9, No. 2, pp. 233-255.

Chinn, C.A. and W.F. Brewer, 2001, "Models of Data: A Theory of How People Evaluate Data", *Cognition and Instruction*, Vol. 19, No. 3, pp. 323-393.

Chuesathuchon, C., 2008, *Computerized Adaptive Testing in Mathematics for Primary Schools in Thailand*, Ph.D., Cowan University Library.

Cicchetti, D.V., 1994, "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology", *Psychological Assessment*, Vol. 6, No. 4, pp. 284–290.

Clark, L.A. and D. Watson, 1995, "Constructing Validity: Basic Issues in Objective Scale Development", *Psychological Assessment*, Vol. 7, No. 3, pp. 309-319.

Cohen, L., L. Manion and K. Morrison, 2018, *Research Methods in Education*, London, Routledge.

Cooper, D.C. and P.S. Schindler, 2001, *Business Research Methods*, McGraw-Hill, New York, London.

Crocker, L. and J. Algina, 2008, *Introduction to Classical and Modern Test Theory*, Cengage Learning, Mason, Ohio.

Cronbach, L.J., 1971, *Objective Structured Clinical Examination (OSCE) with Immediate Feedback in Early (Preclinical) Stages of the Dental Curriculum*, Educational Measurement, Washington, District of Columbia.

Cronbach, L.J. and P.E. Meehl, 1955, "Construct Validity in Psychological Tests", *Psychological Bulletin*, Vol. 52, No. 4, pp. 281-302.

Çıkrıkçı Demirtaşlı, N., 1999, "Psikometride Yeni Ufuklar: Bilgisayar Ortamında Bireye Uyarlanmış Test", *Türk Psikoloji Bülteni*, Vol. 5, No. 13, pp. 31-36.

Dai, S., T.T. Vo, O.J. Kehinde, H. He, Y. Xue, C. Demir and X. Wang, 2021, "Performance of Polytomous IRT Models with Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data", *Frontiers in Education*, Vol. 6, No. 2, pp. 372-390.

Daphine, R., P. Sivakumar and S. Selvakumar, 2020, "A Study on Student's Attitude Towards Online Computer Adaptive Test (CAT) in Physics Education Through Observation Schedule", *Journal of Xidian University*, Vol. 14, No. 5, pp. 4703-4708.

Davey, T., 2011, *A Guide to Computer Adaptive Testing Systems*, Council of Chief State School Officers, Washington, District of Columbia.

Davis, F.D., 1985, *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*, Ph.D. Thesis, Massachusetts Institute of Technology Library.

Davis, F.D., 1989, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", *MIS Quarterly*, Vol. 1, No. 1, pp. 319-340.

Davis, F.D., 1993, "User Acceptance of Information Technology: System Characteristics, User Perceptions and Behavioral Impacts", *International Journal of Man-Machine Studies*, Vol. 38, No. 3, pp. 475-487.

Davis, F.D., R.P. Bagozzi and P.R. Warshaw, 1989, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models", *Management Science,* Vol. 35, No. 8, pp. 982-1003.

Davis, F.D., R.P. Bagozzi and P.R. Warshaw, 1992, "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace 1", *Journal of Applied Social Psychology*, Vol. 22, No. 14, pp. 1111-1132.

De Boeck, P. and M. Wilson, 2004, *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, Springer, Amsterdam, New York.

Deci, E.L., 1971, "Effects of Externally Mediated Rewards on Intri*nsic Motivation",*

Journal of Personality and Social Psychology, Vol. 18, No. 1, pp. 105-125.

DeVellis, R.F., 2003, *Scale Development: Theory and Applications*, Sonic Amateur Games Expo, Washington, District of Columbia.

DeVellis, R.F., 2017, *Scale Development: Theory and Applications*, Sonic Amateur Games Expo, Washington, District of Columbia.

Diagnostic Research, Inc., 1988, "Macintosh or MS-DOS: A Synopsis of What MIS Inanagers and Other Professionals in Fortune 1000 Companies Have to Say", https:// archive.org/ stream/ MacintoshOrMS -DOS/ Macintosh %20or %20MS - DOS _djvu.txt, accessed on January 01, 2024.

Dizon, G., 2016, "Measuring Japanese EFL Student Perceptions of Internet-Based Tests with the Technology Acceptance Model", *Teaching English as a Second Language Electronic Journal(Tesl-Ej)*, Vol. 20, No. 2, pp. 2-17.

Ebel, R.L. and D.A. Frisbie, 1991, *Essentials of Educational Measurement*, Englewood Cliffs, New Jersey, New York.

Eggen, T.J. and A.J. Verschoor, 2006, "Optimal Testing with Easy or Difficult Items in Computerized Adaptive Testing", *Applied Psychological Measurement*, Vol. 30, No. 5, pp. 379-393.

Embretson, S.E. and S.P. Reise, 2000, *Item Response Theory*, Psychology Press, Pennsylvania, New York.

English, R.A., M.D. Reckase and W.M. Patience, 1977, "Application of Tailored Testing to Achievement Measurement", *Behavior Research Methods and Instrumentation*, Vol. 9, No. 2, pp. 158-161.

Eraslan Yalçın, M. and B. Kutlu, 2019, "Examination of Students' Acceptance of and Intention to Use Learning Management Systems Using Extended TAM", *British Journal of Educational Technology*, Vol. 50, No. 5, pp. 2414-2432.

Feketéné Silye, M. and T.B. Wiwczaroski, 2002, "A Critical Review of Selected Computer Assisted Language Testing Instruments", *Acta Agraria Debreceniensis*, Vol. 3, No. 6, pp. 3-6.

Fraenkel, J.R. and N.E. Wallen, 2003, *How to Design and Evaluate Research in Education*, McGraw-Hill Higher Education, Delaware, New York.

French, B.F. and W.H. Finch, 2006, "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance", *Structural Equation Modeling*, Vol. 13, No. 3, pp. 378-402.

Frey, B., 2018, *The SAGE Encyclopedia of Educational Research*, Measurement, and Evaluation, Scientific Advisory Group for Emergencies, Thousand Oaks, California.

García-Fernández, J., A. Postigo, M. Cuesta, C. González-Nuevo, A. Menéndez-Aller and E. García-Cueto, 2022, "To be Direct or not: Reversing Likert Response Format Items", *The Spanish Journal of Psychology*, Vol. 25, No. 2, pp. 24-39.

Ghaderi, M., M. Mogholi and A. Soori, 2014, "Comparing Between Computer Based Tests and Paper-and-Pencil Based Tests", International Journal of Education and Literacy Studies, Vol. 2, No. 4, pp. 36-38.mm Ghaderi, M., M. Mogholi and A. Soori, 2014, "Comparing Between Computer Based Tests and Paper-and-Pencil Based Tests", *International Journal of Education and Literacy Studies*, Vol. 2, No. 4, pp. 36-38.

Glanville, J.L. and T. Wildhagen, 2007, "The Measurement of School Engagement: Assessing Dimensionality and Measurement Invariance Across Race and Ethnicity",

Educational and Psychological Measurement, Vol. 67, No. 6, pp. 1019–1041.

Glas, C.A.W., H. Wainer and E.T. Bradlow, 2003, *MML and EAP estimation in testlet-based adaptive testing*, Kluwer Academic, Dordrecht, New York.

Gliem, J.A. and R.R. Gliem, R.R., 2003, *Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales*, Paper presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Ohio.

Goodwin, N.C., 1987, "Functionality and Usability", *Communications of the ACM*, Vol. 30, No. 3, pp. 229-233.

Gould, J.D., S.J. Boies and C. Lewis, 1991, "Making Usable, Useful, Productivity-Enhancing Computer Applications", *Communications of the ACM*, Vol. 34, No. 1, pp. 74-85.

Hambleton, R.K., 1980, *Test Score Validity and Standard Setting Methods*, in R. A. Berk (Ed.). Criterion-Referenced Measurement: The State of The Art, Johns Hopkins University Press, New York.

Hambleton, R.K., H. Swaminathan, 1984, *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, Boston, Massachusetts.

Hambleton, R.K.,H. Swaminathan and H.J. Rogers, 1991, *Fundamentals of item Response Theory*, Scientific Advisory Group for Emergencies, Newbury Park, California.

Hoffman, K.I. and G.D. Lundberg, 1976, "A Comparison of Computer-Monitored Group Tests with Paper and Pencil Tests", *Educational and Psychological Measurement*, Vol. 36, No. 4, pp. 791-809.

Hong, S., M.L. Malik and M.K. Lee, 2003, "Testing Configural, Metric, Scalar, and Latent Mean Invariance Across Genders in Sociotropy and Autonomy Using a Non-Western Sample", *Educational and Psychological Measurement*, Vol. 63, No. 4, pp. 636–654.

Hu, L. and P.M. Bentler, 1999, "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives", *Structural Equation Modeling*, Vol. 6, No. 1, pp. 1–55.

Karagöz, C., 2022, *Analyzing Teachers' Views toward the Use of EBA during COVID-19 Pandemic Based on the Technology Acceptance Model*, M.S. Thesis, Middle East Technical University.

Klella, A.S., 2020, *The Use of Computers in Testing: Implementation, Attitude, and effectiveness*, Zawiya University Library, Az-Zāwiyah, Libya.

Laerd Statistics, 2018, "Crpnbach's Alpha (α) using SPSS Statistics", https:// statistics. laerd. com/ spss- tutorials/ cronbachs- alpha- using- spss- statistics. php, accessed on October 2, 2023.

Lee, M.K., C.M. Cheung and Z. Chen, 2005, "Acceptance of Internet-Based Learning Medium: The Role of Extrinsic and Intrinsic Motivation", *Information and Management*, Vol. 42, No. 8, pp. 1095-1104.

Lee, Y., K.A. Kozar and K.R. Larsen, 2003, "The Technology Acceptance Model: Past, Present, and Future", *Communications of the Association for information Systems*, Vol. 12, No. 1, pp. 50-76.

Lee, B.C., J.O. Yoon and I. Lee, 2009, "Learners' Acceptance of e-learning in South Korea: Theories and Results", *Computers and Education*, Vol. 53, No. 4, pp. 1320-1329.

Legg, S.M. and D.C. Buhr, 1992, "Computerized Adaptive Testing with Different Groups", *Educational Measurement: Issues and Practice*, Vol. 11, No. 2, pp. 23-27.

Lilley, M., 2007, *The Development and application of Computer Adaptive Testing in a Higher Education Environment*, Ph.D. Thesis, University of Hertfordshire.

Lilley, M., R. Barker and C. Britton, 2004, "The Development and Evaluation of a Software Prototype for Computer-Adaptive Testing", *Computers and Education*, Vol. 43, No. 1, pp. 109-123.

Lilley, M., A. Pyper and P. Wernick, 2011, "Attitudes to and Usage of CAT in Assessment in Higher Education", *Innovation in Teaching and Learning in Information and Computer Sciences*, Vol. 10, No. 3, pp. 28-37.

Lim, L., 2020, "Development and initial validation of the Computer-Delivered Test Acceptance Questionnaire for Secondary and High School Students", *Journal of Psychoeducational Assessment*, Vol. 38, No.2, pp. 182-194.

Linn, R.L., 1993, *Educational Measurement* (3th ed), American Council on Education and the Oryx Press, New York.

Lord, F.M., 1952, "A Theory of Test Scores", *Psychometric Monographs*, Vol. 7, No. 1, pp. 1-84.

Lord, F.M. and M.L. Stocking, 1988, Item response theory. In J.P. Keeves (Eds.), *Educational Research, Methodology, and Measurement: An International Handbook*, Methodology of Educational Measurement and Assessment, Pergamon, New York.

Lucke, J.F. 2005a, "The α and the ω of Congeneric test theory: An Extension of Reliability and Internal Consistency to Heterogeneous Tests", *Applied Psychological Measurement*, VOl. 29, No. 1, pp. 65-81.

Lucke, J.F. 2005b, "Rassling the hog: The Influence of Correlated Item Error on Internal Consistency", Classical Reliability, and Congeneric Reliability", *Applied Psychological Measurement*, Vol. 29, No. 1, pp. 106–125.

Madsen, H., 1986, "Evaluating a Computer-Adaptive ESL Placement Test", *CALICO Journal*, Vol. 41, No. 50, pp.41-50.

Marsh, H.W. and D. Hocevar, 1985, "Application of Confirmatory Factor Analysis to the Study of Self-Concept: First-and Higher Order Factor Models and Their Invariance Across Groups", *Psychological Bulletin*, Vol. 97, No. 3, pp. 562-580.

Martin, J.T., and J.R. McBride, 1983, *Reliability and Validity of Adaptive Ability Tests in a Military Setting*, New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing, New York.

Martin, J.T., J.R. McBride and D.J. Weiss, 1983, *Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population*, Minnesota University Minneapolis Computerized Adaptive Testing, Minnesota.

Martínez-Torres, M.D. R., S.L. Toral Marín, F.B. García, S.G. Vazquez, M.A. Oliva and T. Torres, 2008, "A Technological Acceptance of e-learning Tools Used in Practical and Laboratory Teaching, According to the European Higher Education Area", *Behaviour and Information Technology*, Vol. 27, No. 6, pp. 495-505.

McBride, J.R., 2001, *The Marine Corps Exploratory Development Project*: 1977-1982. In W. A. Sands, B.K. Waters and J R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*, Psychological Association, Washington, District of Columbia.

McDonald, A.S.,2002, "The Impact of Individual Differences on the Equivalence of Computer-Based and Paper-and-Pencil Educational Assessments", *Computers and*

Education, Vol. 39, No. 3, pp. 299-312.

McGlohen, M. and H.H Chang, 2008, "Combining Computer Adaptive Testing Technology with Cognitively Diagnostic Assessment", *Behavior Research Methods*, Vol. 40, No. 1, pp. 808-821.

McNeish, D., 2018, "Thanks coefficient alpha, we'll take it from here", *Psychological Methods*, Vol. 23, No. 3, pp. 412–433.

Meade, A.W., E.C. Johnson and P.W. Braddy, 2008, "Power and Sensitivity of Alternative Fit Indices in Tests of Measurement Invariance", *Journal of Applied Psychology*, Vol. 93, No. 3, pp. 568-586.

Moe, K.C. And M.F. Johnson, 1988, "Participants' Reactions to Computerized Testing", *Journal of Educational Computing Research*, Vol. 4, No. 1, pp. 79-86.

Moreno, K.E., C.D. Wetzel, J.R. McBride and D.J. Weiss, 1984, "Relationship Between Corresponding Armed Services Vocational Aptitude Battery and Computerized Adaptive Testing Subtests", *Applied Psychological Measurement*, Vol. 8, No. 2, pp. 155-163.

Mun, Y.Y. and Y. Hwang, 2003, "Predicting the Use of Web-Based Information Systems: Self-Efficacy, Enjoyment, Learning Goal Orientation, and the Technology Acceptance Model", *International Journal of Human-Computer Studi*es, Vol. 59, No. 4, pp. 431-449.

Muraki, E., 1992, "A Generalized Partial Credit Model: Application of an EM Algorithm", *Applied Psychological Measurement*, Vol. 16, No.1, pp. 159-176.

Muthén, L.K. and B.O. Muthén, 2012, *Mplus: Statistical Analysis with Latent Variables; User's Guide*, Muthén et Muthén, Los Angeles, California.

Nering, M.L. and R. Ostini, 2010, *Handbook of Polytomous Item Response Theory Models*, Routledge, Amsterdam, New York.

Nering, M.L. and R. Ostini, 2011, *Handbook of Polytomous Item Response Theory Models*, Taylor and Francis, Vol. 2, No. 1, pp. 1-20.

Nikolaus, S., C. Bode, E. Taal, H.E. Vonkeman, C.A. Glas, and M.A. van de Laar, 2014, "Acceptance of New Technology: A Usability Test of a Computerized Adaptive Test for Fatigue in Rheumatoid Arthritis", *JMIR Human Factors*, Vol. 1, No. 1, pp. 567-587.

Nugroho, R.A., N.S. Kusumawati and O.C. Ambarwati, 2018, "Students Perception on the Use of Computer Based Test", *In IOP Conference Series: Materials Science and Engineering*, Vol. 306, No. 1, pp. 012103-012119.

Ogilvie, R.W., T.C. Trusk and A.V. Blue, 1999, "Students' Attitudes towards Computer Testing in a Basic Science Course", *Medical Education*, Vol. 33, No. 11, pp. 828-831.

Organization for Economic Cooperation and Development, 2010, "Computer-Based Assessmnet of Student Skills in Science", https://www. oecd- ilibrary. org/ education/ pisa- computer- based- assessment-of - student- skills- in- science/ features-of- the- computer- based- items- and- performance _9789264082038 -7 -en, accessed on August 19, 2023.

Özbaşı, D., 2016, "Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamasına ve Kağıt- Kalem Testine Katılan Öğrenci Görüşlerinin İncelenmesi", *Journal of International Social Researc*h, Vol. 9, No. 42, pp.123-145.

Papanastasiou, E., 2003, *Computer-Adaptive Testing in Science Education. In Proceedings of the 6th International Conference on Computer Based Learning in Science*,

Nicosia, Cyprus, Vol. 1, No. 1, pp. 965-971.

Pino-Silva, J., 2008, "Student Perceptions of Computerized Tests", *ELT Journal*, Vol. 62, No. 2, pp. 148-156.

Putnick, D.L. and M.H. Bornstein, 2016, "Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research", *Developmental Review*, Vol. 41, pp. 71-90.

R Core Team 2014, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rasch, G., 1960, "Probabilistic Models for Some Intelligence and Attainment Tests", *Copenhagen: Denish Institute for Educational Research*, Vol.2, No.3, pp. 158-178.

Ratna, P.A. And S. Mehra, 2015, "Exploring the Acceptance for e–learning Using Technology Acceptance Model among University Students in India", *International Journal of Process Management and Benchmarking*, Vol. 5, No. 2, pp. 194-210.

Reise, S.P. and D.A. Revicki, 2015, *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, Routledge, New York.

Reise, S.P., K.F. Widaman and R.H. Pugh, 1993, "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance", *Psychological Bulletin*, Vol. 114, No. 3, pp. 552-566.

Rosenberg, B.D. and M.A. Navarro, 2018, Semantic Differential scaling. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, Scientific Advisory Group for Emergencies, Los Angeles.

Rovinelli R.J. and R.K. Hambleton, 1977, "On The Use of Content Specialists in the

Assessment of Criterion-Referenced Test Item Validity", *Dutch Journal of Educational Research*, Vol. 2, No. 1, pp. 49-60.

RStudio, Inc, 2015, *RStudio: Integrated Development Environment*, American Journal of Plant Sciences, Boston, Massachusetts.

Rutkowski, L. and D. Svetina, 2014, ""Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys", *Educational and Psychological Measurement*, Vol. 74, No. 1, pp. 31-57.

Saadé, R. and B. Bahli, 2005, "The Impact of Cognitive Absorption on Perceived Usefulness and Perceived Ease of Use in on-line Learning: An Extension of the Technology Acceptance Model", *Information and Managemen*t, Vol. 42, No. 2, pp. 317-327.

Salkind N.J., 2010, *Encyclopedia of Research Design, Sonic Amateur Games Expo* Scientific Advisory Group for Emergencies, Washington, District of Columbia.

Samejima, F., 1969, "Estimation of Latent Ability Using a Response Pattern of Graded Scores", *Psychometr Monogr Suppl*, Vol. 34, No. 4, pp. 100-141.

Samsudin, M.A., N. Norfarah and T. Chut, 2020, "Psychometric Assessment of Mathematics Algebra Item Banks for Computerized Adaptive Test", *Technology Reports of Kansai University*, Vol, 62, No. 1, pp. 2061-2076.

Schleicher, A., 2019, *PISA 2018: Insights and Interpretations*, The Organisation for Economic Co-operation and Development, André Pascal.

Schmidt, F.L., V.W. Urry and J.F. Gugel, 1978, "Computer Assisted Tailored Testing: Examinee Reactions and Evaluations", *Educational and Psychological Measurement*, Vol. 38, No. 2, pp. 265-273.

Segall, D.O. And K.E. Moreno, 1999, "Development of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery", *Innovations in Computerized Assessment*, Vol. 1, No. 1, pp. 35-65.

Senese, P.V., M.H. Bornstein, O.M. Haynes, G. Rossi and P.A. Venuti, 2012, "Cross-Cultural Comparison of Mothers' Beliefs about Their Parenting Very Young Children", *Infant Behavior and Development*, Vol. 35, No. 1, pp. 479–488.

Smith, B. and P. Caputi, 2004, "The Development of the Attitude Towards Computerized Assessment Scale", *Journal of Educational Computing Research*, Vol. 31, No. 4, pp. 407-422.

Stensen, K. and S. Lydersen, 2022, "Internal Consistency: From Alpha to Omega", *Tidsskrift for den Norske Laegeforening: Tidsskrift for Praktisk Medicin*, Vol. 142, No. 12, pp. 142-167.

Student Selection and Placement Center, 2022, "Statistics of 2022 YKS", https: //www. osym. gov. tr/ TR, 23867/ 2022- ykssinav- sonuclarina- iliskin- sayisal- bilgiler. ht, accessed on October 12, 2024.

Stricker, L.J., G.Z. Wilder and D.A. Rock, 2004, "Attitudes about the Computer-Based Test of English as a Foreign Language", *Computers in Human Behavior*, Vol. 20, No. 1, pp. 37-54.

Şad, S.N., A. Kış, M. Demir and N. Özer, 2016, "Meta-Analysis of the Relationship Between Mathematics Anxiety and Mathematics Achievement", *Pegem Journal of Education and Instruction*, Vol. 6, No. 3, pp. 371-392.

Şanlı, R., 2003, *Students' Perceptions about Online Assessment: A Case Study*, M.S. Thesis, Middle East Technical University.

Teo, T., C.B. Lee and C.S. Chai, 2008, "Understanding Pre-Service Teachers' Computer Attitudes: Applying and Extending the Technology Acceptance Model", *Journal of Computer Assisted Learning*, Vol. 24, No. 2, pp. 128-143.

Teo, T. and J. Noyes, 2011, "An Assessment of the Influence of Perceived Enjoyment and Attitude on the Intention to Use Technology Among Pre-Service Teachers: A Structural Equation Modeling Approach", *Computers and Education*, Vol. 57, No. 2, pp. 1645-1653.

Terzis, V. and A.A. Economides, 2011a, "The Acceptance and Use of Computer Based Assessment", *Computers and Education*, Vol. 56, No. 4, pp. 1032-1044.

Terzis, V. and A.A. Economides, 2011b, "Computer Based Assessment: Gender Differences in Perceptions and Acceptance", *Computers in Human Behavior*, Vol. 27, No. 6, pp. 2108-2122.

Thorndike, R.M., G.K. Cunningham, R.L. Thorndike and E.P. Hagen, 1991, *Assessing the Structural and Concurrent Validity of a Shortened Version of the Domestic Violence Healthcare Providers' Survey Questionnaire for Use in Sweden*, Measurement and evaluation in psychology and education, New York.

Tonidandel, S. and M.A. Quinones, 2000, "Psychological Reactions to Adaptive Testing", *International Journal of Selection and Assessment*, Vol. 8, No. 1, pp. 7-15.

Turkish Ministry of Education, PISA 2022 "Turkey Report ", https:// pisa. meb. gov. tr/meb _iys _dosyalar /2023 _12/ 05125555 _pisa2022 _rapor _051223. pdf, accessed on October 30, 2024.

Ullman, J.B., 2001, Structural Equation Modeling. In B. Tabachnick and L. S. Fidell (Eds.), *Using Multivariate Statistics*, Allyn and Bacon, Boston.

Ullman, J.B., 2006, "Structural Equation Modeling: Reviewing the Basics and Moving Forward", *Journal of Personality Assessment*, Vol. 87, No. 1, pp. 35-50.

Van der Linden, W.J. and G.A. W. Glas, 2000, *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic, Boston.

Van der Linden, W.J. and P.J. Pashley, 2009, *Item Selection and Ability Estimation in Adaptive Testing*, In Elements of Adaptive Testing, Springer New York.

Venkatesh, V., 2000, "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model", *Information Systems Research*, Vol. 11, No. 4, pp. 342–365.

Vicino, F.L. and K.E Moreno, 1997, Human Factors in the CAT System: A Pilot Study. In W.A. Sands, B.K. Waters and J.R. McBride.(Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*, Psychological Association, Washington, District of Columbia.

Vispoel, W.P., T.R. Rocklin and T. Wang, 1994, "Individual Differences and Test Administration Procedures: A Comparison of Fixed-item, Computerized-Adaptive, and Self-Adapted Testing", *Applied Measurement in Education*, Vol. 7, No. 1, pp. 53-79.

Wainer, D.J., D.J. Dorans, R. Flaugher, B.F. Green, L. Mislevy, L. Steinberg and D. Thissen, 1990, *Computerized Adaptive Testing: A Primer, Lawrence Erlbaum Associates*, Hillsdale, New Jersey.

Webster, E.J., 1989, *Plnyfzilness and Coirzpziters at Work. Ph. D. Thesis, Stern School of Business Administration*, New York University Library, New York.

Weiss, D.J., 1985, "Adaptive Testing by Computer", *Journal of Consulting and Clinical*

Psychology, Vol. 53, No. 6, pp. 774.

Weiss, D.J., 2004, "Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education", *Measurement and Evaluation in Counseling and Development*, Vol. 37, No. 2, pp. 70-84.

Weiss, D.J. and G.G. Kingsbury, 1984, "Application of Computerized Adaptive Testing to Educational Problems", *Journal of Educational Measurement*, Vol. 21, No. 4, pp. 361-375.

Widaman, K.F. and S.P. Reise, 1997, Exploring the Measurement Invariance of Psychological Instruments: Applications in the Substance Use Domain. In K.J. Bryant, M.E. Windle, and S.G. West (Eds.), *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, Psychological Association, Washington.

Wright, B.D. and M.H. Stone, 1979, *Best Test Design: Rasch Measurement*, Illinois Mesa Press, Chicago.

Yamamoto, K., H.J. Shin and L. Khorramdel, 2019, *Introduction of Multistage Adaptive Testing Design in PISA 2018*, Organisation for Economic Co-operation and Development, André Pascal.

Yang Y. and S.B. Green, 2011, "Coefficient Alpha: A Reliability Coefficient for the 21st Century?", *Journal of Psychoeducational Assessment*, Vol. 29, No. 4, pp. 377-392.

Yen W.M., 1984, "Effects of Local Item Dependence On the Fit and Equating Performance of the Three-Parameter Logistic Model", *Applied Psychological Measurement*, Vol. 8, pp. 125-145.

Yenilmez, K., 2007, "İlköğretim Öğrencilerinin Matematik Dersine Yönelik Tutumları",

Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, Vol. 23, No. 23, pp. 51-59.

Zilles, C., R.T Deloatch, J. Bailey, B.B. Khattar, W. Fagen, C. Heeren, D. Mussulman and M. West, 2015, "Computerized Testing: A Vision and Initial Experiences", *ASEE Annual Conference and Exposition*, Vol. 1, No. 1, pp. 26-387.

Zinbarg, R.E., W. Revelle, I. Yovel and W. Li, 2005, "Cronbach's α, Revelle's β and Mcdonald's ωH: Their Relations With Each Other and Two Alternative Conceptualizatıons of Reliability", *Psychometrika*, Vol. 70, No. 1, pp. 1–11.

APPENDIX A: EXAMINEES' ATTITUDES TOWARD COMPUTERIZED ADAPTIVE TEST SCALE (PILOT STUDY)

Bilgisayar ortamında bireye uyarlanmış testlere karşı tutum ölçeği

Merhaba!

Bu ölçek senin az önce yanıtlamış olduğun teste karşı tutumunu belirlemeyi amaçlıyor.

Lütfen her bir ifade için sana en uygun olan seçeneğin altına X işareti koy.

Table A.1. Bilgisayar ortamında bireye uyarlanmış testlere karşı tutum ölçeği.

Maddeler		**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Katılıyorum**	**Kesinlikle Katılıyorum**
1.	Ekranda işaretleme yapmak yorucuydu.				
2.	Biraz önce yanıtladığım testte, diğer sınavlarıma göre daha az soruyla karşılaştım.				
3.	Eğlenceli bir testti.				
4.	Testi çözmeyi sevdim.				
5.	Bilgisayar kullanımı ile ilgili öğretmenimden yardım istedim.				

Table A.1. Bilgisayar ortamında bireye uyarlanmış testler. (cont.)

Maddeler		**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Katılıyorum**	**Kesinlikle Katılıyorum**
6.	Biraz önceki testi çözmek yerine, başka bir şey yapmak isterdim.				
7.	Biraz önce yanıtladığım test, diğer sınavlarıma göre daha az zamanımı aldı.				
8.	Soruları ekrandan okumakta zorlandım.				
9.	Bilgisayardan okumak, soruları anlamamı zorlaştırdı.				
10.	Testin hemen bitmesini istedim.				
11.	Benim bilgi seviyeme daha uygun sorularla karşılaştım.				
12.	Soruları bilgisayardan okuyup, işlemleri kağıtta yapmak beni yordu.				

Table A.1. Bilgisayar ortamında bireye uyarlanmış testler. (cont.)

Maddeler		**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Katılıyorum**	**Kesinlikle Katılıyorum**
13	Soruların boş bırakılamaması beni her soruyu düşünmeye yönlendirdi.				
14.	Testi çözmek zevkliydi.				
15.	Aşırı zor veya aşırı kolay sorularla karşılaşmadım.				
					Teşekkürler

APPENDIX B: EXAMINEES' ATTITUDES TOWARD COMPUTERIZED ADAPTIVE TEST SCALE (FINAL STUDY)

Bilgisayar Ortamında Bireye Uyarlanmış Testlere Karşı Tutum Ölçeği

Cinsiyet: Kız / Erkek

Merhaba!

Bu ölçek senin az önce yanıtlamış olduğun teste karşı tutumunu belirlemeyi amaçlıyor.

Lütfen her bir ifade için sana en uygun olan seçeneğin altına X işareti koy.

Teşekkürler

Table B.1. Bilgisayar ortamında bireye uyarlanmış testlere karşı tutum ölçeği.

Maddeler		**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Katılıyorum**	**Kesinlikle Katılıyorum**
1.	Ekranda işaretleme yapmak yorucuydu.				
2.	Biraz önce yanıtladığım testte, diğer sınavlarıma göre daha az soruyla karşılaştım.				
3.	Eğlenceli bir testti.				
4.	Testi çözmeyi sevdim.				
5.	Biraz önceki testi çözmek yerine, başka bir şey yapmak isterdim.				

Table B.1. Bilgisayar ortamında bireye uyarlanmış testler. (cont.)

Maddeler		**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Katılıyorum**	**Kesinlikle Katılıyorum**
6.	Biraz önce yanıtladığım test, diğer sınavlarıma göre daha az zamanımı aldı.				
7.	Soruları ekrandan okumakta zorlandım.				
8.	Bilgisayardan okumak, soruları anlamamı zorlaştırdı.				
9.	Benim bilgi seviyeme daha uygun sorularla karşılaştım.				
10.	Soruları bilgisayardan okuyup, işlemleri kağıtta yapmak beni yordu.				
11.	Testi çözmek zevkliydi.				
12.	Aşırı zor veya aşırı kolay sorularla karşılaşmadım.				

APPENDIX C: YEN'S Q3 STATISTICS' RESULTS 1

Yen's Q3 Statistics of Pilot Study are shown below.

Table C.1. Yen's Q3 statistics' results

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
M1	1.000	0.151	0.042	-0.098	-0.031	-0.089	0.070	0.119
M2	0.151	1.000	0.056	-0.128	0.038	-0.140	0.358	0.069
M3	0.042	0.056	1.000	-0.219	-0.051	-0.281	-0.038	-0.107
M4	-0.098	-0.128	-0.219	1.000	-0.045	-0.163	-0.064	-0.110
M5	-0.031	0.038	-0.051	-0.045	1.000	-0.053	-0.005	0.198
M6	-0.089	-0.140	-0.281	-0.163	-0.053	1.000	-0.213	-0.081
M7	0.070	0.358	-0.038	-0.064	-0.005	-0.213	1.000	0.036
M8	0.119	0.069	-0.107	-0.110	0.198	-0.081	0.036	1.000
M9	0.102	0.084	-0.130	-0.280	0.167	-0.038	0.065	0.512
M10	-0.062	-0.120	-0.245	-0.161	0.074	0.142	-0.056	-0.086
M11	-0.051	0.042	-0.137	-0.128	0.092	-0.130	0.106	0.074
M12	0.160	-0.003	-0.174	-0.093	0.075	0.020	0.099	0.196
M13	-0.043	0.104	-0.048	0.094	0.018	0.000	0.048	0.020
M14	-0.092	0.004	-0.279	-0.241	-0.023	-0.201	0.037	-0.068
M15	0.060	0.039	-0.119	-0.140	0.134	-0.056	0.097	0.031

Table C.2. Yen's Q3 statistics' results (pilot study) 1.

	M9	**M10**	**M11**	**M12**	**M13**	**M14**	**M15**
M1	0.102	-0.062	-0.051	0.160	-0.043	-0.092	0.060
M2	0.084	-0.120	0.042	-0.003	0.104	-0.004	0.039
M3	-0.130	-0.245	-0.137	-0.174	-0.048	-0.279	-0.119
M4	-0.280	-0.161	-0.128	-0.093	0.094	-0.241	-0.140
M5	0.167	0.074	0.092	0.075	0.018	-0.023	0.134
M6	-0.038	0.142	-0.130	0.020	0.000	-0.201	-0.056
M7	0.065	-0.056	0.106	0.099	0.048	0.037	0.097
M8	0.512	-0.086	0.074	0.196	0.020	-0.068	0.031
M9	1.000	-0.024	0.082	0.288	-0.064	-0.054	0.076
M10	-0.024	1.000	-0.028	0.000	0.036	-0.198	0.087
M11	0.082	-0.028	1.000	0.004	0.128	-0.010	0.146
M12	0.288	0.000	0.004	1.000	-0.115	-0.122	0.036
M13	-0.064	0.036	0.128	-0.115	1.000	-0.045	-0.007
M14	-0.054	-0.198	-0.010	-0.122	-0.045	1.000	0.074
M15	0.076	0.087	0.146	0.036	-0.007	0.074	1.000

APPENDIX D: YEN'S Q3 STATISTICS' RESULTS 2

Yen's Q3 Statistics of Final Study are shown below.

Table D.1. Yen's Q3 statistics' results

	M1	**M2**	**M3**	**M4**	**M5**	**M6**
M1	1.000	0.024	-0.076	-0.058	0.070	0.120
M2	0.024	1.000	-0.074	0.026	-0.024	0.138
M3	-0.076	0.074	1.000	-0.023	-0.192	-0.068
M4	-0.058	0.026	-0.023	1.000	-0.030	-0.152
M5	0.070	-0.024	-0.192	-0.030	1.000	-0.148
M6	0.120	0.138	-0.068	-0.152	-0.148	1.000
M7	0.125	0.013	-0.242	-0.210	-0.059	0.111
M8	0.080	-0.073	-0.336	-0.092	0.040	-0.020
M9	-0.077	0.011	-0.087	-0.249	-0.073	0.006
M10	0.143	-0.158	-0.281	-0.302	0.075	0.009
M11	-0.180	-0.191	-0.242	-0.383	-0.132	-0.102
M12	-0.084	0.026	-0.097	-0.061	-0.224	0.044

Table D.2. Yen's Q3 statistics' results (final study) 1.

	M7	**M8**	**M9**	**M10**	**M11**	**M12**
M1	0.125	0.080	-0.077	0.143	-0.180	-0.084
M2	0.013	-0.073	0.011	-0.158	-0.191	0.026
M3	-0.242	-0.336	-0.087	-0.281	-0.242	-0.097
M4	-0.210	-0.092	-0.249	-0.302	-0.383	-0.061
M5	-0.059	0.040	-0.073	0.075	-0.132	-0.224
M6	0.111	-0.020	0.006	-0.009	-0.102	0.044
M7	1.000	0.362	-0.009	0.173	-0.069	-0.110
M8	0.362	1.000	-0.052	0.222	-0.078	-0.139
M9	-0.009	-0.052	1.000	-0.008	-0.054	-0.038
M10	0.173	0.222	-0.008	1.000	0.135	-0.067
M11	0.069	-0.078	-0.054	0.135	1.000	-0.002
M12	0.110	-0.139	-0.038	-0.067	-0.002	1.000

APPENDIX E: ITEM CHARACTERISTIC CURVES 1

ICCs of Pilot Study are shown below.

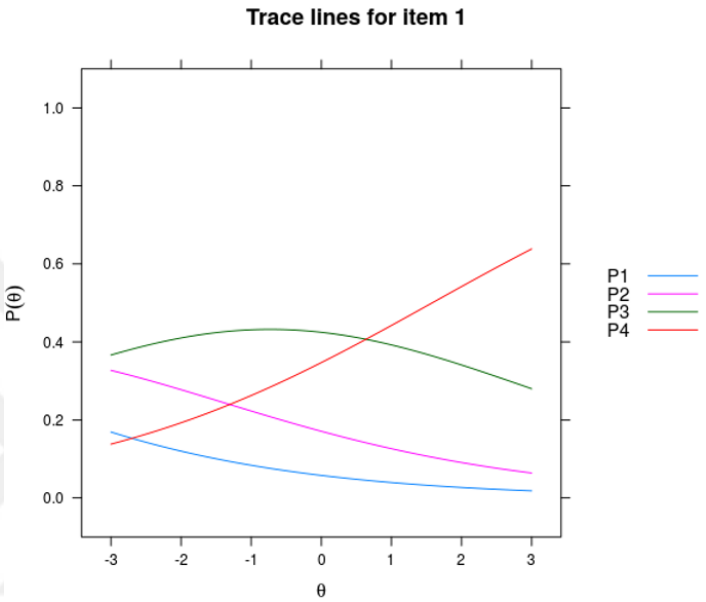

Figure E.1. Item characteristic curves 1 for pilot study.

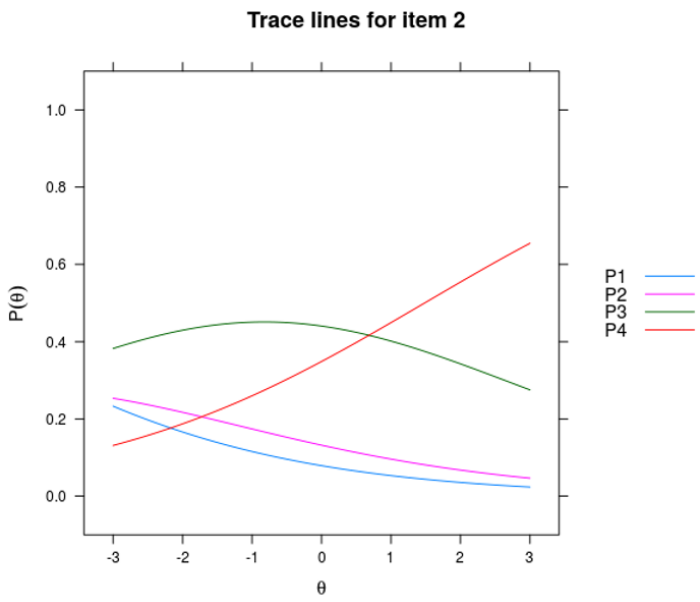

Figure E.2. Item characteristic curves 2 for pilot study.

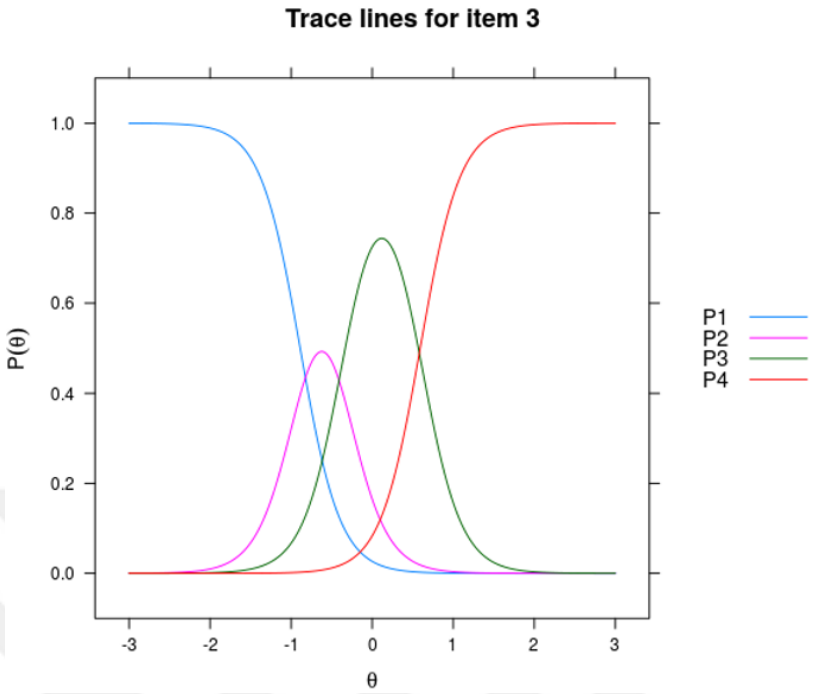

Figure E.3. Item characteristic curves 3 for pilot study.

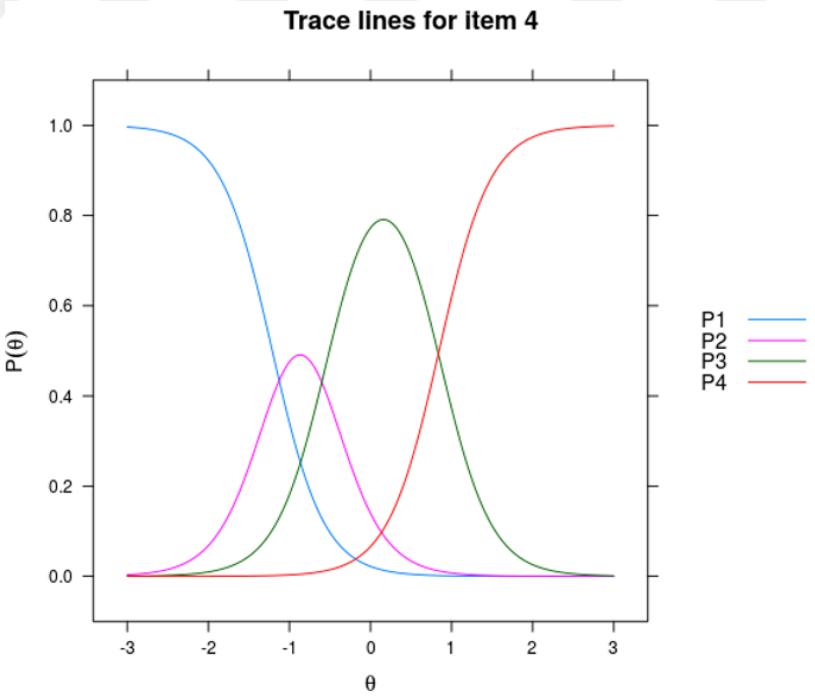

Figure E.4. Item characteristic curves 4 for pilot study.

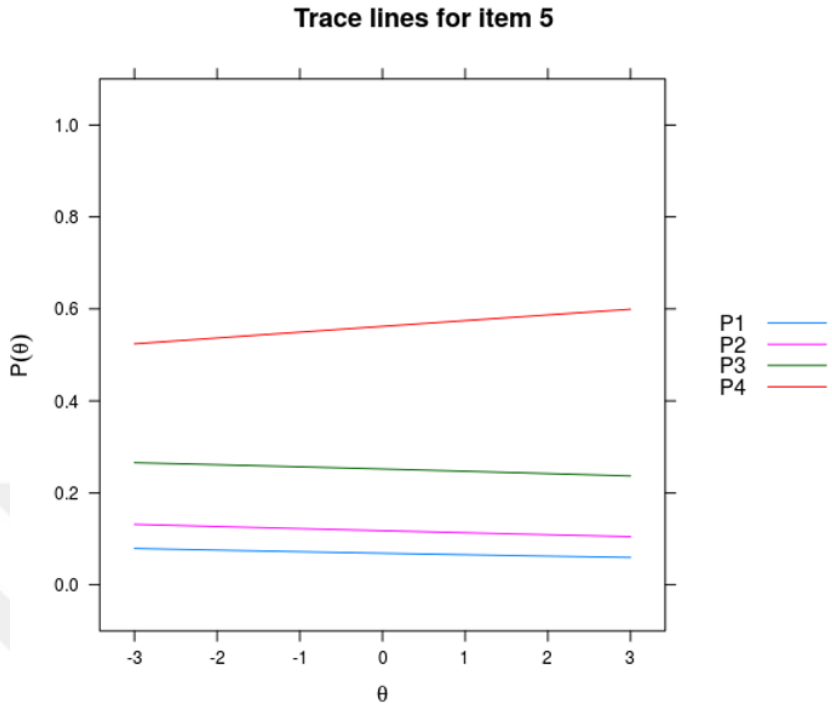

Figure E.5. Item characteristic curves 5 for pilot study.

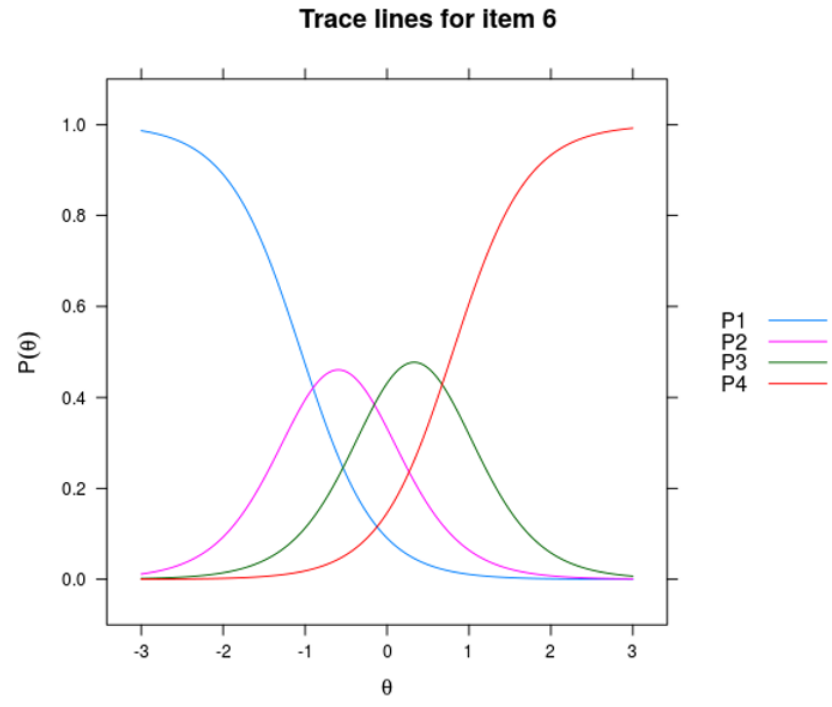

Figure E.6. Item characteristic curves 6 for pilot study.

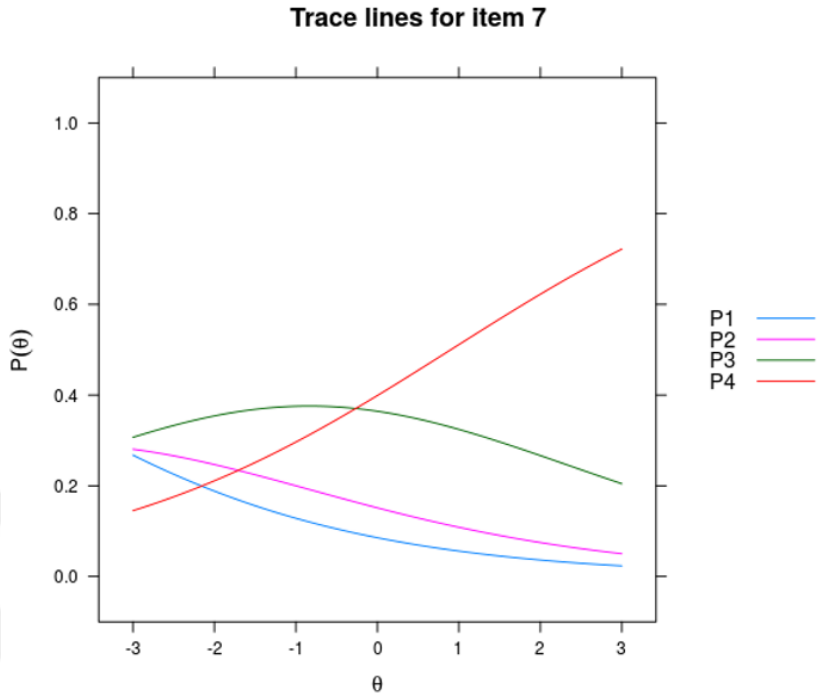

Figure E.7. Item characteristic curves 7 for pilot study.

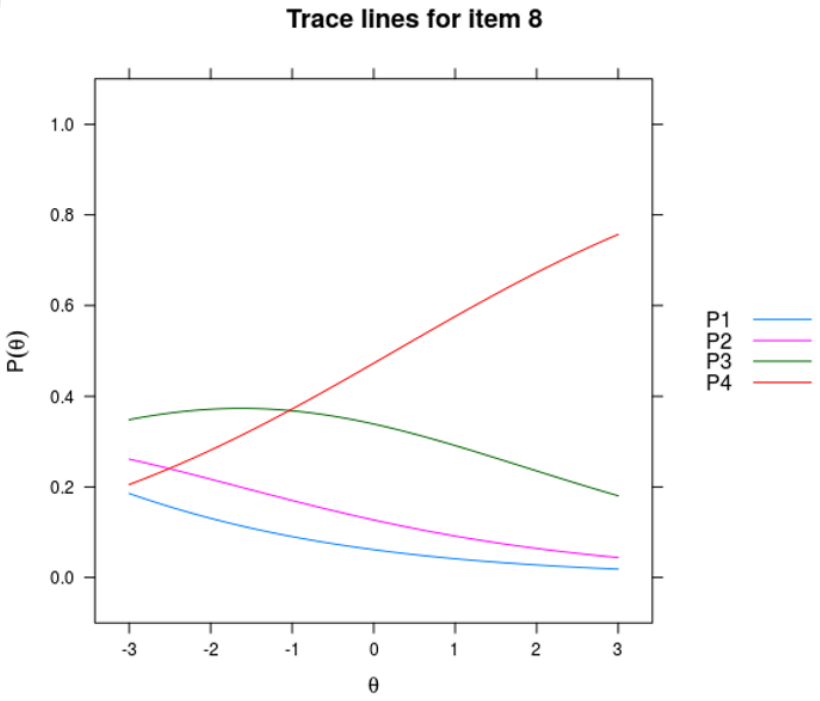

Figure E.8. Item characteristic curves 8 for pilot study.

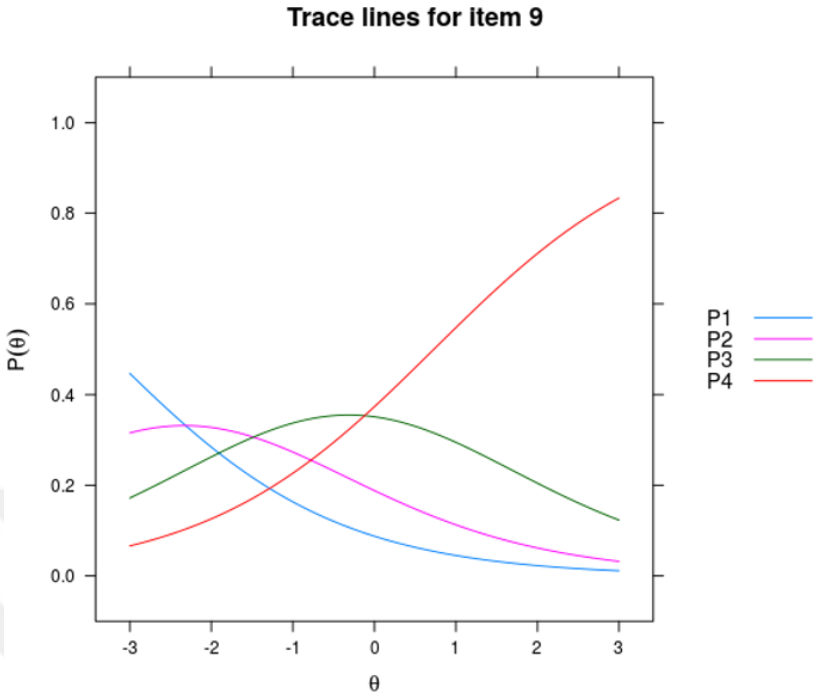

Figure E.9. Item characteristic curves 9 for pilot study.

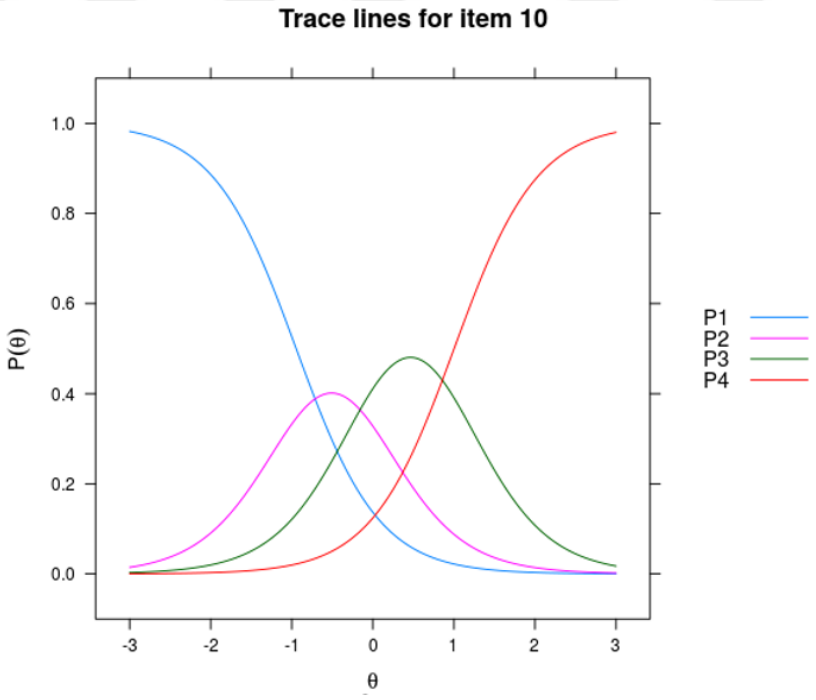

Figure E.10. Item characteristic curves 10 for pilot study.

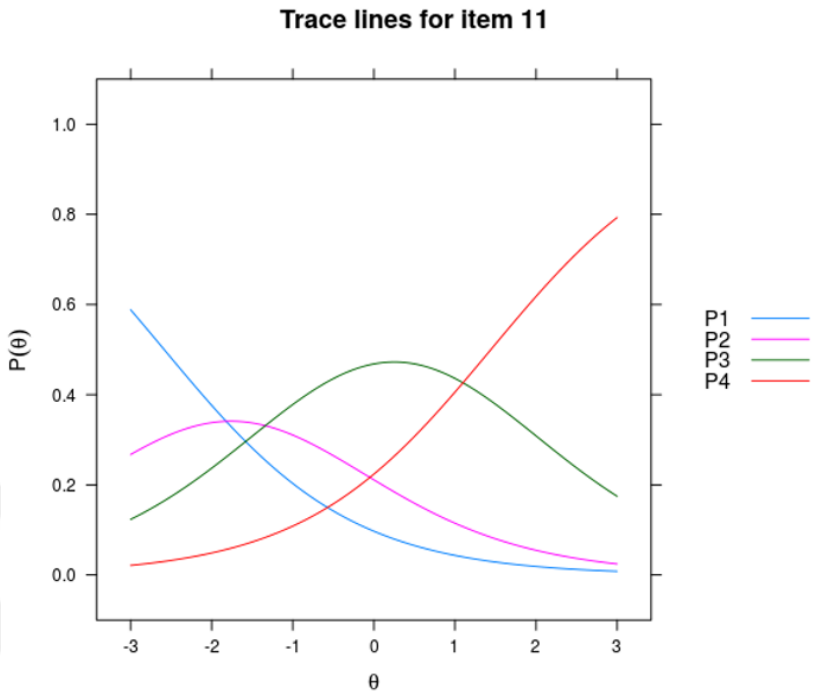

Figure E.11. Item characteristic curves 11 for pilot study.

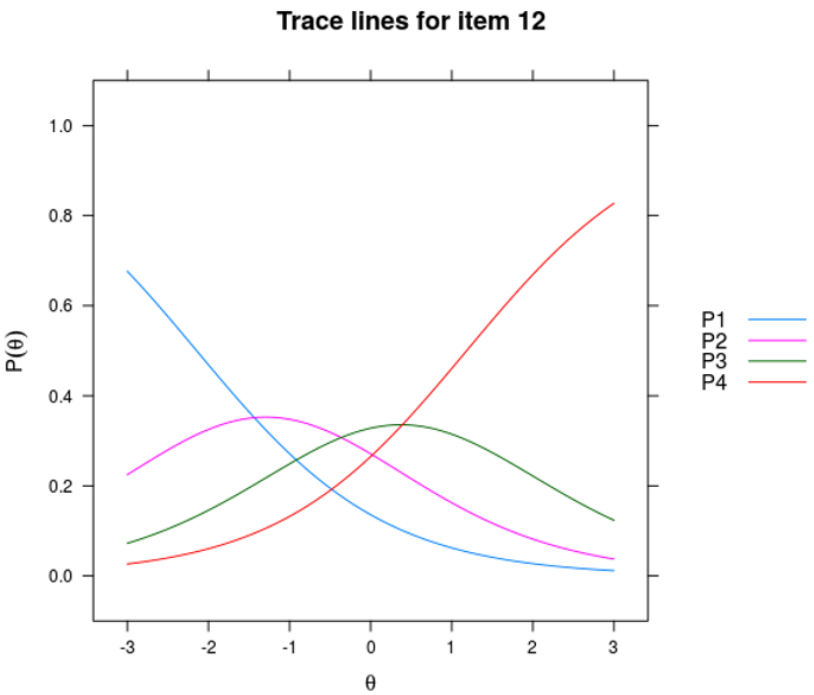

Figure E.12. Item characteristic curves 12 for pilot study.

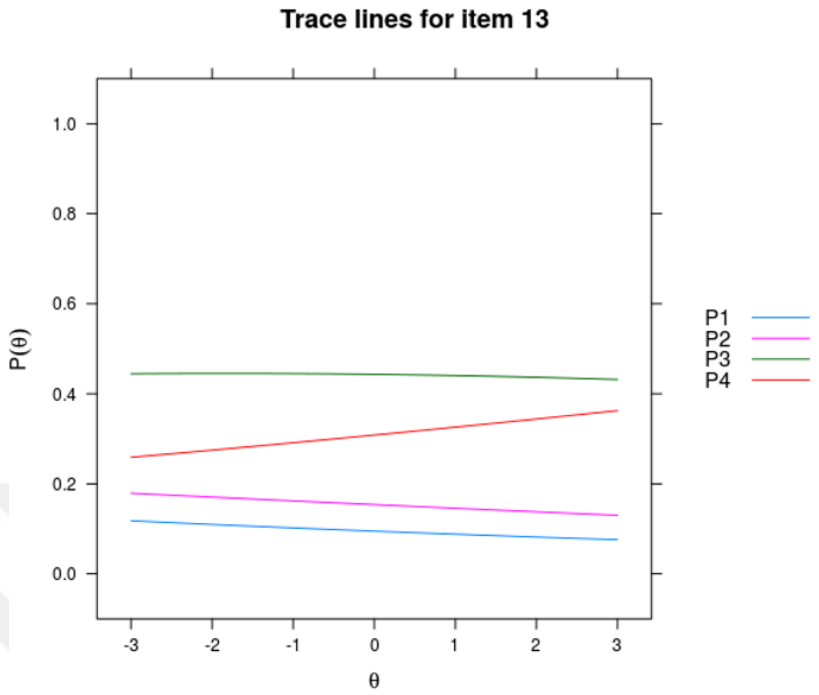

Figure E.13. Item characteristic curves 13 for pilot study.

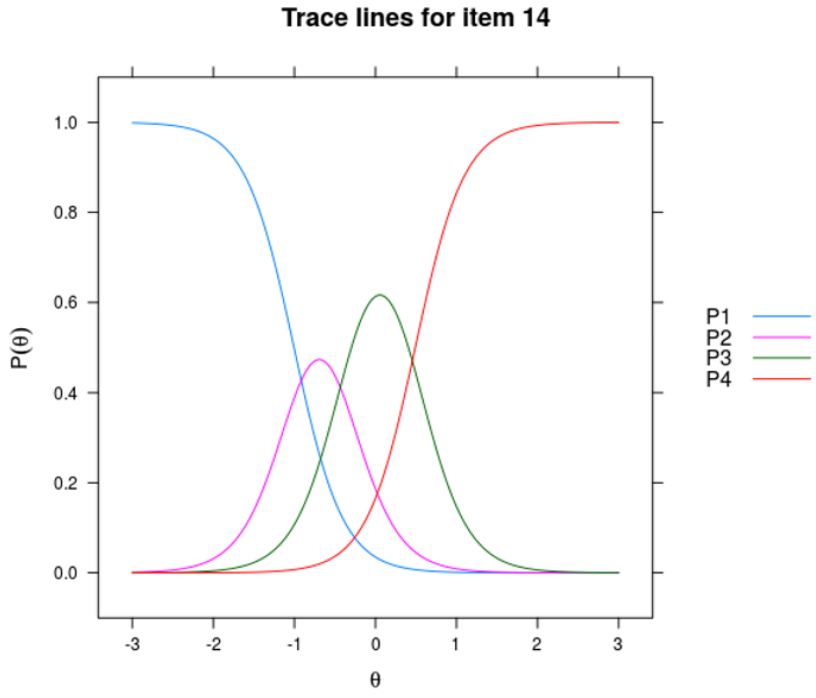

Figure E.14. Item characteristic curves 14 for pilot study.

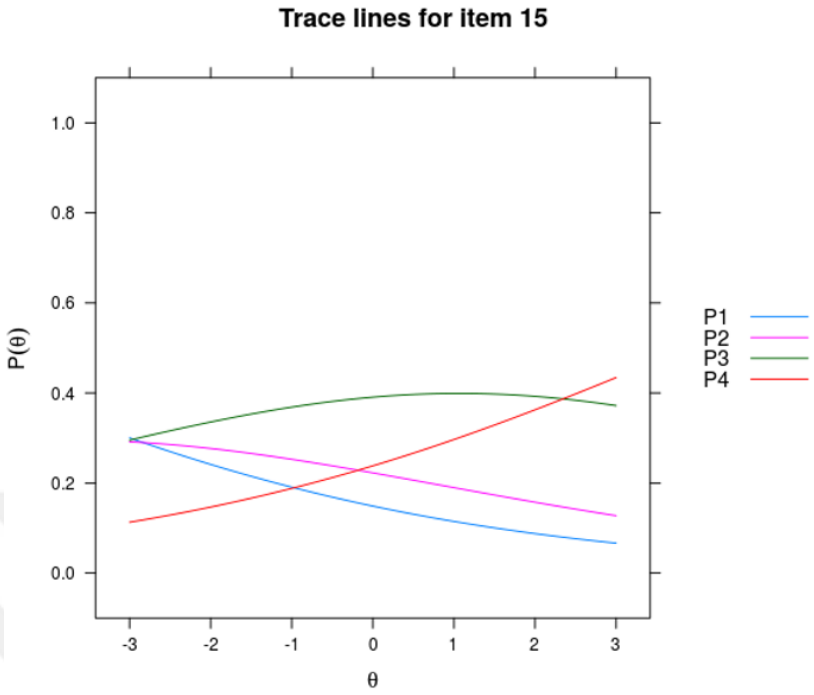

Figure E.15. Item characteristic curves 15 for pilot study.

APPENDIX F: ITEM CHARACTERISTIC CURVES 2

ICCs of Final Study are shown below.

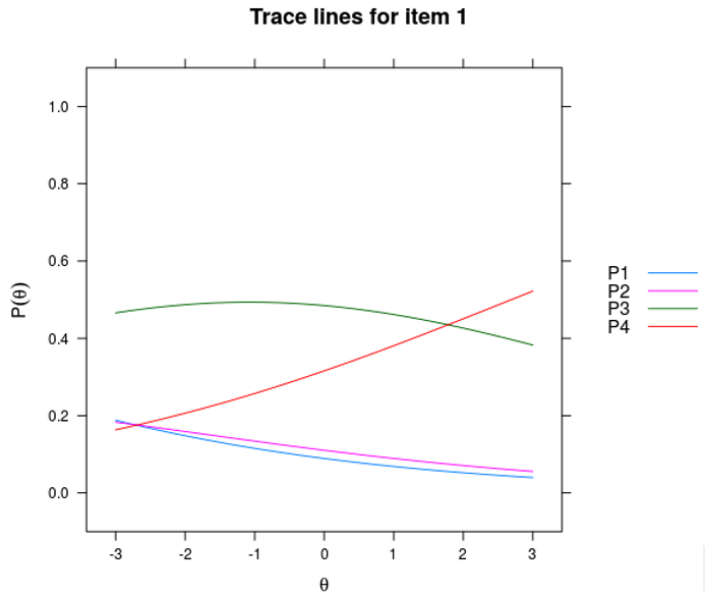

Figure F.1. Item characteristic curves 1 for final study.

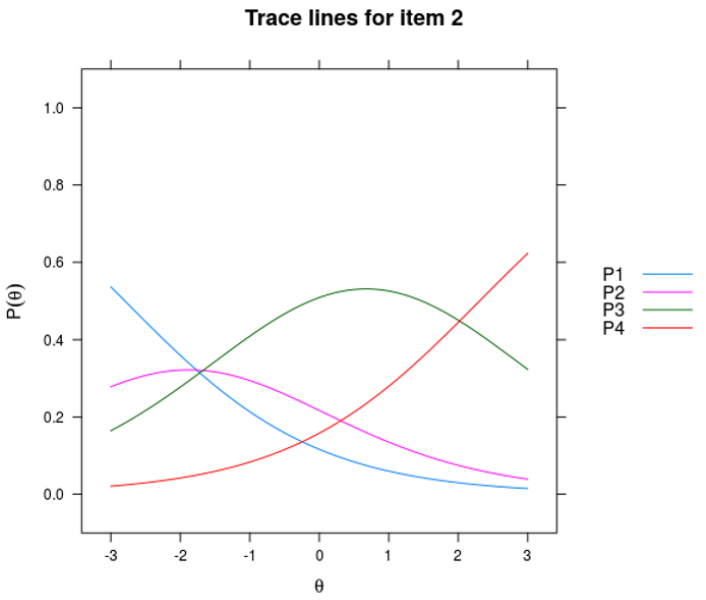

Figure F.2. Item characteristic curves 2 for final study.

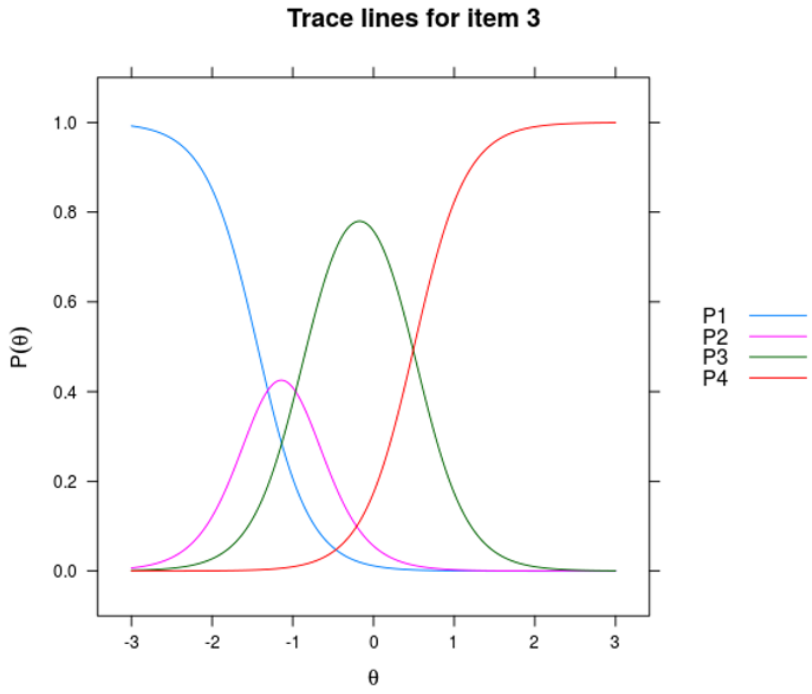

Figure F.3. Item characteristic curves 3 for final study.

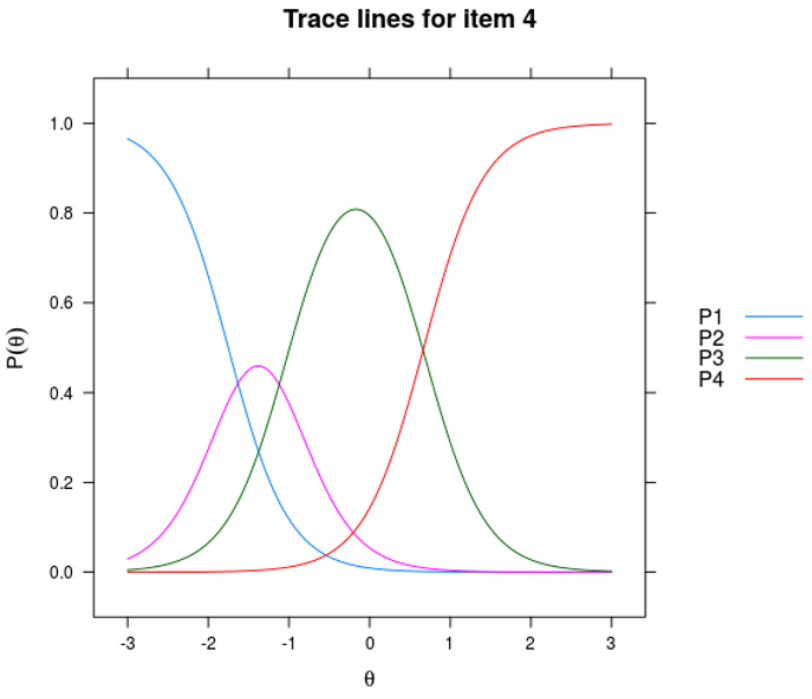

Figure F.4. Item characteristic curves 4 for final study.

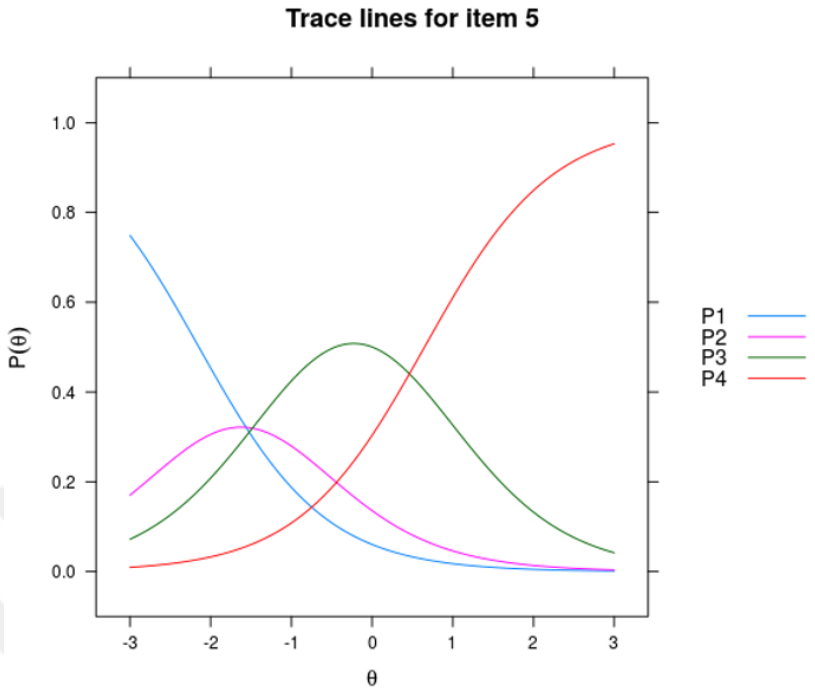

Figure F.5. Item characteristic curves 5 for final study.

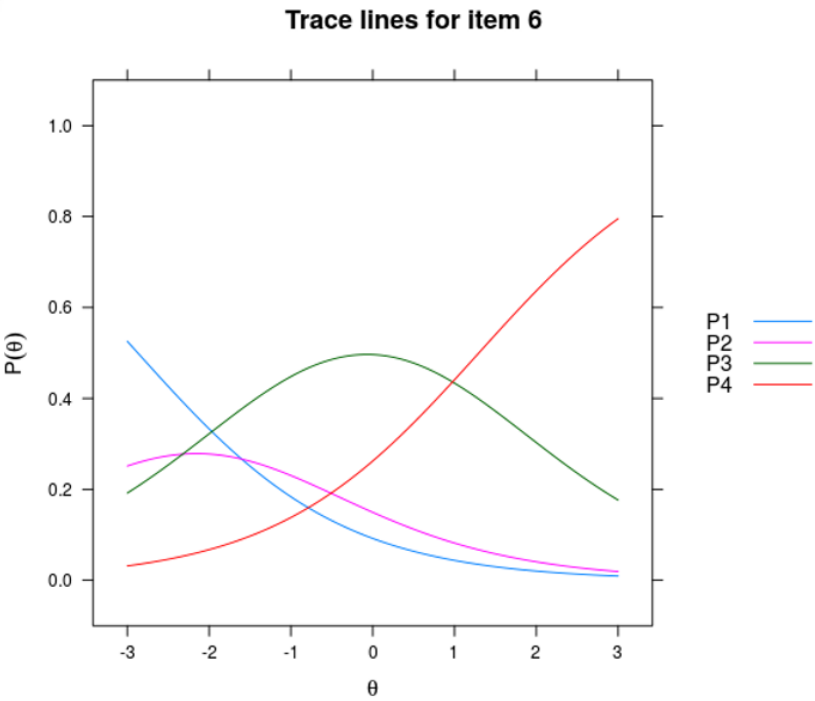

Figure F.6. Item characteristic curves 6 for final study.

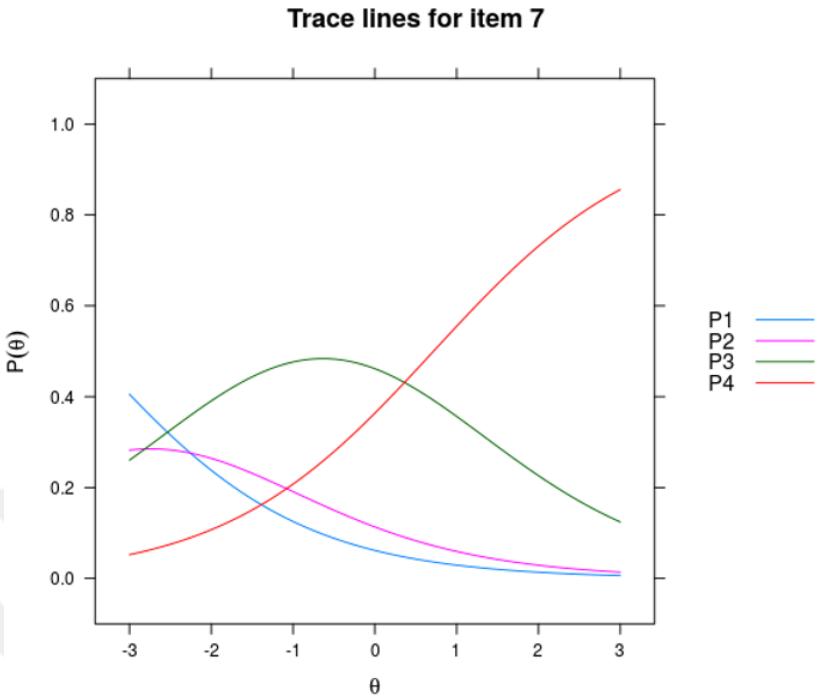

Figure F.7. Item characteristic curves 7 for final study.

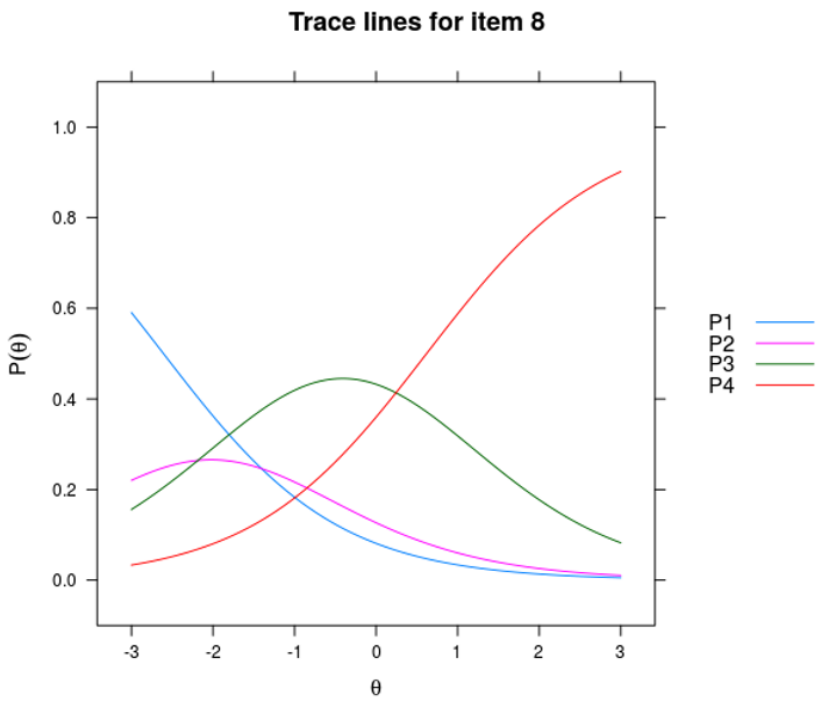

Figure F.8. Item characteristic curves 8 for final study.

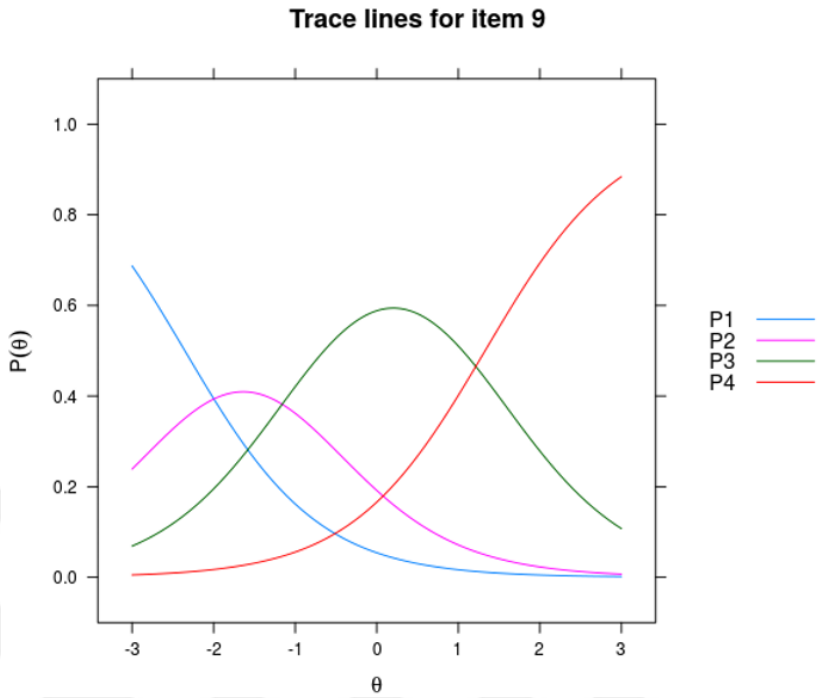

Figure F.9. Item characteristic curves 9 for final study.

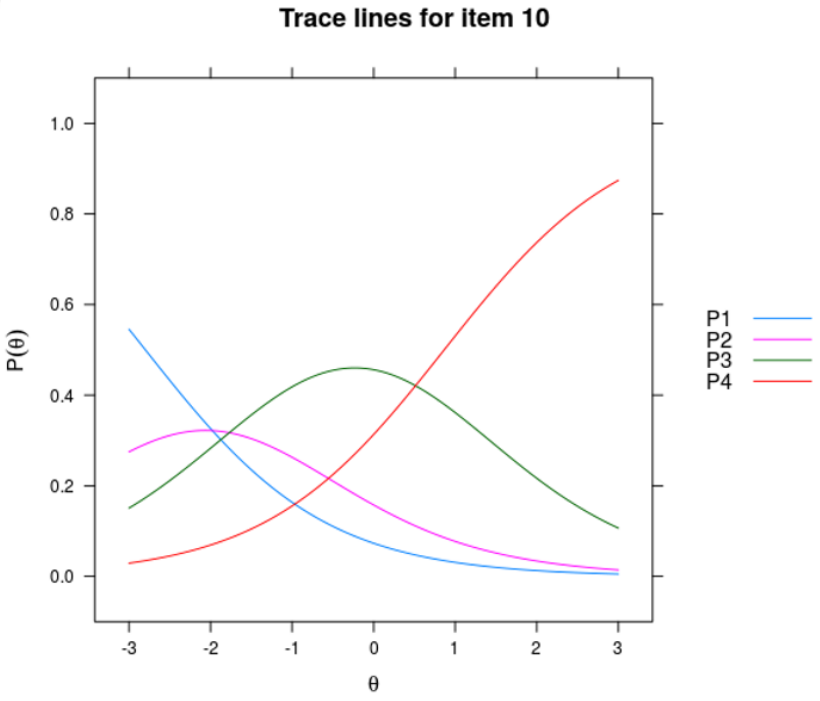

Figure F.10. Item characteristic curves 10 for final study.

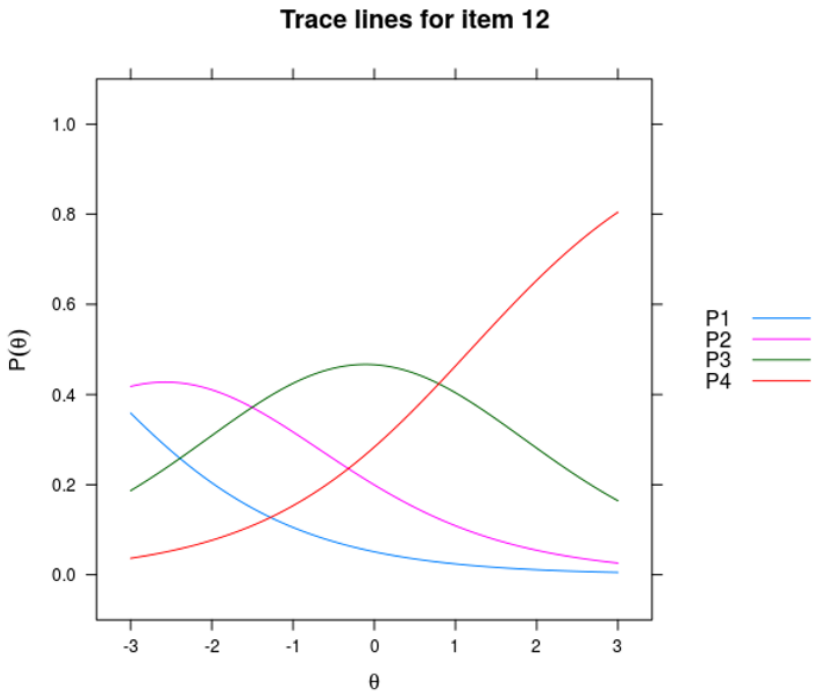

Figure F.11. Item characteristic curves 11 for final study.

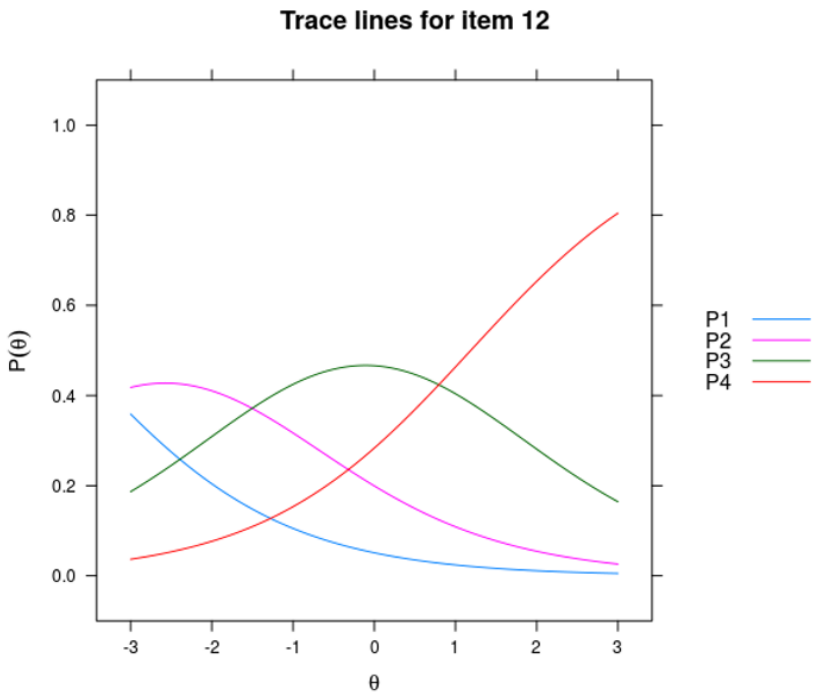

Figure F.12. Item characteristic curves 12 for final study.