

DESIGNING A MULTISTAGE TEST FOR ASSESSMENT LITERACY

by

Belgin Eriz

B.S., Primary Mathematics Education, Boğaziçi University, 2022

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 Mathematics and Science Education
Boğaziçi University

2025

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my advisor, Assoc. Prof. Serkan Arıkan, for his invaluable guidance and unwavering support throughout this journey. He was actively involved at every stage of the research process, offering insightful feedback and consistent encouragement. His mentorship not only deepened my understanding of educational assessment, but I also learned a great deal from him about self-discipline, academic integrity, and the importance of collaboration. Regardless of the career path I follow, he will always remain one of my most respected role models. I consider myself truly fortunate to have worked under his supervision.

I would also like to express my sincere appreciation to my thesis committee members, Assoc. Prof. Beyza Aksu Dünya and Assist. Prof. Osman Akşit, for their generous time and constructive feedback, which greatly enriched this study. I am also thankful to Assoc. Prof. Eren Can Aybek, whose expertise and guidance in the multistage test development process were instrumental to the success of this research.

I gratefully acknowledge the financial support provided by TÜBİTAK through the 1002-A Program. My genuine thanks also go to the faculty members and participants from various universities who supported the data collection process. Their cooperation was crucial to the successful completion of this study. I am also appreciative of the members of the BounAdaptiveTestLab for their collaborative spirit and contributions. Their support offered both practical help and meaningful intellectual engagement.

I am deeply thankful to my parents, Cevat and Nilgün Eriz. The strength and grace with which they faced life's challenges taught me how to persevere without losing hope or joy. Their unwavering love, patience, and belief in me gave me the courage to pursue my dreams. I am also profoundly grateful to my dear sisters, Nurtaç, Rumeysa, and Esma Eriz, for their constant love and support. I truly believe that children who grow up surrounded by warmth and kindness begin life one step ahead-and I was

fortunate to be one of them. In moments of doubt, their words and presence uplifted me and reminded me of the strength that comes from family.

I feel incredibly fortunate to have had the steady support of my friends throughout this journey. I am especially grateful to Evra Konuk and Fatma Soyucok, whose presence and encouragement continually reminded me of my purpose and helped me find direction whenever I felt lost. Thank you for always being there when I needed a walk, a coffee chat, or a motivational talk.

I would also like to express my heartfelt thanks to Sümeysra Yıldız, İdil Yüksektepe, and Nurefşan Alım. Since our undergraduate years, you witnessed so much of my struggle from surviving Math111 to navigating this entire journey. I am grateful for your constant compassion and empowerment.

I am truly thankful to Betül Yıldırım, Züleyha Terzi, and Beyzanur Yalvaç. Our shared reflections on the joys and frustrations of academia, and the friendship we built along the way were a constant source of laughter, comfort, and motivation.

Lastly, I wish to express my deepest gratitude to my grandmother, Fatma, whose spirit has been the foundation of my character and the quiet strength behind all that I strive to be. She taught me that unconditional love is real and that love has the power to transform everything it touches. Although she could not witness the end of this journey, her smile and prayers remain with me every step of the way. The support and love she gave throughout her life continue to sustain me, more than enough to last a lifetime.

ABSTRACT

DESIGNING A MULTISTAGE TEST FOR ASSESSMENT LITERACY

Assessment literacy is a critical component of effective teaching and learning, as it underpins the validity, reliability, and fairness of educational assessments. Measuring the assessment literacy of in-service and pre-service teachers is essential for understanding their knowledge, skills, and competencies in implementing sound assessment practices. However, existing theoretical frameworks and assessment tools often fail to capture the multifaceted nature of this construct. This study aims to bridge this gap by extending a widely used assessment literacy framework into an assessment framework and developing a multistage test for measuring assessment literacy efficiently. In the first phase, insights from 19 expert interviews and an extensive literature review were used to construct an 8-dimensional assessment framework for assessment literacy. The extended framework incorporated two novel dimensions: competency in technology-based assessment and competency in developing assessment tools. A 105-item pool was then developed based on this framework and administered via paper-and-pencil format to 1,223 pre-service teachers from 11 universities to estimate item parameters. A 1-3-3 panel multistage test was developed on the Concerto platform, based on Item Response Theory (IRT) analysis of pilot test findings. Each module in the MST included one item per dimension and met targeted psychometric criteria. The MST was subsequently administered to 106 pre-service teachers. IRT-based analysis confirmed the effectiveness of the MST in measuring assessment literacy across a range of ability levels, with a decreasing standard error across stages, demonstrating both efficiency and precision.

ÖZET

ÖLÇME-DEĞERLENDİRME OKURYAZARLIĞINA YÖNELİK ÇOK AŞAMALI BİR TEST GELİŞTİRME ÇALIŞMASI

Ölçme-değerlendirme okuryazarlığı, eğitimde yapılan değerlendirmelerin geçerliği, güvenirliği ve adilliği açısından temel teşkil ettiğinden, etkili öğretim ve öğrenmenin kritik bir bileşenidir. Ancak, mevcut kuramsal çerçeveler ve ölçme araçları, bu yapının çok yönlü doğasını ele almakta çoğunlukla yetersiz kalmaktadır. Bu çalışma, yaygın olarak kullanılan bir ölçme-değerlendirme okuryazarlığı çerçevesini bir değerlendirme çerçevesi olarak genişletmeyi ve ölçme-değerlendirme okuryazarlığını etkili biçimde ölçecek çok aşamalı bir test geliştirmeyi amaçlamaktadır. İlk aşamada, 19 uzmanla yapılan görüşmeler ve kapsamlı bir alan yazını taraması doğrultusunda bir ölçme-değerlendirme okuryazarlığı kuramsal çerçevesi, sekiz boyutlu bir değerlendirme çerçevesine genişletmiştir. Bu çerçeveye dayalı olarak geliştirilen 105 maddelik madde havuzu, 11 üniversiteden 1.223 öğretmen adayına kâğıt-kalem formatında uygulanmıştır. Pilot uygulamadan elde edilen veriler Madde Tepki Kuramı (MTK) ile analiz edilerek Concerto platformunda 1-3-3 panelli çok aşamalı bireye uyarlanmış test tasarlanmıştır. Testte yer alan her modül, her bir boyutu temsil eden bir madde içermekte ve hedeflenen psikometrik ölçütleri karşılamaktadır. Geliştirilen çok aşamalı bireye uyarlanmış test, 106 öğretmen adayına uygulanmıştır. MTK tabanlı analizler sonucunda, testin farklı yetenek düzeylerinde değerlendirme okuryazarlığını etkili biçimde ölçtüğü ve test aşamaları ilerledikçe standart hatada azalma sağladığı görülmüştür. Bu bulgular, geliştirilen çok aşamalı bireye uyarlanmış testin hem etkin hem de güvenilir bir ölçme aracı olduğunu ortaya koymuştur.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	v
ÖZET	vi
LIST OF FIGURES	ix
LIST OF TABLES	xi
LIST OF SYMBOLS	xv
LIST OF ACRONYMS/ABBREVIATIONS	xvi
1. INTRODUCTION	1
1.1. The Background of Study	1
1.2. The Definition of Assessment Literacy	2
1.3. The Multistage Testing	4
1.4. The Significance of the Study	9
1.5. The Purposes of the Study	10
1.6. Research Questions	10
2. LITERATURE REVIEW	11
2.1. Existing Frameworks for Assessment Literacy	11
2.2. Review of Global Research on Measurement Instruments for AL	18
2.3. Review of Research on Measurement Instruments for AL in Turkey	22
3. METHODOLOGY	26
3.1. Participants	26
3.2. Procedure	29
3.2.1. The Development of an Assessment Framework for Assessment- Literacy	29
3.2.2. The Test Development for Measuring Assessment Literacy	31
3.2.2.1. Pilot testing	31
3.2.2.2. The Development of Multistage Test	35
3.3. Data Analysis	39
3.3.1. Reliability	39

3.3.2. Validity	41
3.3.3. Item Response Theory	44
4. RESULTS	49
4.1. Results of the Assessment Framework Study	49
4.2. Results of Pilot Testing	57
4.2.1. The Reliability of Pilot Test Results	57
4.2.2. The Validity of Pilot Test Results	58
4.2.3. Results of Item Response Theory Analysis	60
4.3. Results of the MST	69
4.3.1. The Reliability of MST Results	79
5. DISCUSSION	81
5.1. Discussion of RQ1: Core Dimensions of Assessment Literacy	81
5.2. Discussion of RQ2: Reliability and Efficiency of the MST	84
5.3. The Promotion of Assessment Literacy	87
6. LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH	89
6.1. Limitations	89
6.2. Suggestions for Further Research	90
REFERENCES	91
APPENDIX A: IOC INDICES	104
APPENDIX B: EFA RESULTS	108
APPENDIX C: CFA MODEL ERROR	132

LIST OF FIGURES

Figure 1.1.	An example of 1-2 MST design.	6
Figure 3.1.	Phases of the assessment literacy framework development.	31
Figure 3.2.	Sample item 1.	36
Figure 3.3.	Sample item 2.	36
Figure 3.4.	Sample item 3.	37
Figure 3.5.	Three-stage 1-3-3 MST design.	37
Figure 3.6.	A typical item characteristic curve.	45
Figure 4.1.	Dimensions of the assessment framework for assessment literacy. .	53
Figure 4.2.	Item characteristic curves for pilot test items.	61
Figure 4.3.	Histogram of item difficulty parameters.	67
Figure 4.4.	Test information function and SEM across ability levels.	68
Figure 4.5.	Item information functions for pilot test items.	69
Figure 4.6.	Distribution of routing module selections by participants.	72
Figure 4.7.	The TIF for stage 1.	77

Figure 4.8.	The TIF for stage 2.	78
Figure 4.9.	The TIF for stage 3.	78
Figure 4.10.	The TIFs for all modules.	79
Figure 4.11.	Standard errors of ability estimation across stages.	80
Figure C.1.	CFA model error.	132

LIST OF TABLES

Table 1.1.	The summary of ability estimation methods.	7
Table 3.1.	Demographic characteristics of experts.	26
Table 3.2.	Demographic characteristics of participants for pilot.	28
Table 3.3.	Demographic characteristics of participants for MST.	29
Table 3.4.	MMS design structure.	34
Table 3.5.	MMS design for IOC calculations.	43
Table 4.1.	Teacher assessment literacy framework.	53
Table 4.2.	Cronbach's alpha and McDonald's omega estimates for each booklet.	57
Table 4.3.	Table of specification.	58
Table 4.4.	Summary of EFA results across seven booklets.	60
Table 4.5.	Chi-square item fit values for pilot test.	62
Table 4.6.	Descriptive statistics of the item difficulty parameters.	66
Table 4.7.	Distribution of item difficulty parameters for MST modules.	70
Table 4.8.	Ability estimates and standard errors in MST.	72

Table 4.9.	Marginal reliability for each module in the MST.	79
Table A.1.	IOC indices.	104
Table B.1.	KMO and Bartlett's test for booklet 1.	108
Table B.2.	Communalities for booklet 1.	108
Table B.3.	Variance explained and factor eigenvalues for booklet 1.	109
Table B.4.	Factor matrix for booklet 1.	110
Table B.5.	Pattern matrix for booklet 1.	110
Table B.6.	KMO and Bartlett's test for booklet 2.	110
Table B.7.	Communalities for booklet 2.	111
Table B.8.	Variance explained and factor eigenvalues for booklet 2.	112
Table B.9.	Factor matrix for booklet 2.	113
Table B.10.	KMO and Bartlett's test for booklet 3.	113
Table B.11.	Communalities for booklet 3.	114
Table B.12.	Variance explained and factor eigenvalues for booklet 3.	115
Table B.13.	Factor matrix for booklet 3.	116
Table B.14.	Pattern matrix for booklet 3.	116

Table B.15. KMO and Bartlett's test for booklet 4.	117
Table B.16. Communalities for booklet 4.	118
Table B.17. Variance explained and factor eigenvalues for booklet 4.	119
Table B.18. Factor matrix for booklet 4.	120
Table B.19. Pattern matrix for booklet 4.	120
Table B.20. KMO and Bartlett's test for booklet 5.	121
Table B.21. Communalities for booklet 4.	122
Table B.22. Variance explained and factor eigenvalues for booklet 5.	123
Table B.23. Factor matrix for booklet 5.	124
Table B.24. Pattern matrix for booklet 5.	124
Table B.25. KMO and Bartlett's test for booklet 6.	125
Table B.26. Communalities for booklet 6.	125
Table B.27. Variance explained and factor eigenvalues for booklet 6.	126
Table B.28. Factor matrix for booklet 6.	127
Table B.29. Pattern matrix for booklet 6.	128
Table B.30. KMO and Bartlett's test for booklet 7.	128

Table B.31. Communalities for booklet 7.	129
Table B.32. Variance explained and factor eigenvalues for booklet 7.	129
Table B.33. Factor matrix for booklet 7.	130
Table B.34. Pattern matrix for booklet 6.	131



LIST OF SYMBOLS

$E[SE(\theta^2)]$	Average squared standard error of ability estimates
$S - \chi^2$	Orlando and Thissen's chi-square item fit statistic
$V(O)$	Variance of the observed sum
$V(Xi)$	Variance of responses to item i
α	Cronbach's alpha
μ	Judges' average rating of item i on all objectives
μ_k	Judges' average rating of item i on the kth objective
θ	Ability estimate
ω	McDonald's omega

LIST OF ACRONYMS/ABBREVIATIONS

1-PL	One-Parameter Logistic
2-PL	Two-Parameter Logistic
3-PL	Three-Parameter Logistic
ACAI	Approaches to Classroom Assessment Inventory
ALI	Assessment Literacy Inventory
AL	Assessment Literacy
CAT	Computerized Adaptive Test
CALI	Classroom Assessment Inventory
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CTT	Classical Test Theory
DPI	Defined Population Interval Metho
EAP	Expected a Posteriori
EFA	Exploratory Factor Analysis
IB	International Baccalaureate
ICC	Item Characteristic Curve
IIF	Item Information Function
IOC	Item-Objective Congruence
IRT	Item Response Theory
KMO	Kaiser-Meyer-Olkin
KR-20	Kuder-Richardson 20
MAP	Maximum a Posteriori
MLE	Maximum Likelihood Estimation
MMS	Multiple Matrix Sampling
MoNE	Ministry of National Education
MST	Multistage Test
PE	Physical Education
PPT	Paper-and-Pencil Test

RMSEA	Root Mean Square Error of Approximation
SEM	Standard Error of Measurement
TALIP	Teacher Assessment Literacy in Practice
TALQ	Teacher Assessment Literacy Questionnaire
TIF	Test Information Function
TIMSS	Trends in International Mathematics and Science Study
TLI	Tucker-Lewis Index



1. INTRODUCTION

This chapter provides the background of the study, defines assessment literacy, identifies multistage testing, outlines the significance of the study, and presents the research purposes and research questions.

1.1. The Background of Study

Assessment is recognized as a fundamental principle that underpins the curriculum in numerous educational systems globally, and it also plays a critical role in teacher professionalism and the quality of teaching (Brookhart, 2011; McMillan, 2013; Phelps, 2015; Shepard, 2000). Regardless of grade level or subject area, assessment skills are essential for educators, as they rely on assessment data to make important instructional and evaluative decisions about their students. Teachers are expected to monitor students' progress, evaluate their learning outcomes, analyze the data obtained from several sources systematically, and adjust the instruction in response to these results. It is estimated that activities linked to assessments take up to 50 percent of a teacher's time (Stiggins, 1991). Numerous research has indicated that effective assessment can enhance teaching and learning by promoting student engagement, motivation, and achievement (Brookhart, 2011). However, several studies highlight that teachers frequently feel inadequately equipped to carry out educational assessments at a professional standard due to inadequate training in their pre-service programs, leading them to encounter challenges in making assessment-related decisions (Kauffman, 2022; Lumadi, 2013). Therefore, thorough exploration of assessment knowledge and skills that teachers are expected to possess becomes a critical demand. At this point, the notion of assessment literacy, which was first coined by Stiggins (1991), acts as an umbrella term covering a broad range of educational assessment practices. It is essential for teachers to possess assessment literacy which facilitates the use of student learning information for effective teaching and addressing students' learning needs (Pastore and Andrade, 2019). Thus, measuring and promoting teachers' assessment

literacy has been a considerable focus over the last two decades (DeLuca *et al.*, 2016).

1.2. The Definition of Assessment Literacy

Assessment literacy (AL) is a multifaceted concept that encompasses various understandings and abilities, lacking a globally accepted formal definition. The concepts of "assessment literacy," "assessment knowledge," and "assessment competence" have been used interchangeably in previous studies (Koh *et al.*, 2017), yet the specific term of "assessment literacy" was initially introduced by Stiggins (1991) to refer to the basic understanding of educational assessment and the related skills needed to apply such knowledge to various measures of student achievement. Over time, the understanding of AL has been evolving responding to advancements and changing demands in the field. However, researchers typically define it as a blend of assessment-related knowledge and skills, aligning with Stiggins's definition.

Such an approach can be attributed to The Standards for Teacher Competence in Educational Assessment of Students, commonly referred to as the 1990 Standards, that were developed to address the problem of inadequate assessment training for teachers (American Federation of Teachers *et al.*, 1990). According to the Standards, classroom teachers are expected to demonstrate proficiency in: Choosing and Developing Assessment Methods; Administering, Scoring, and Interpreting Assessment Results; Using Assessment Results for Decision Making and Grading; Communicating Assessment Results; and Recognizing Unethical Assessment Practices. The extent of an educator's AL is clearly defined and outlined by these Standards through measurable competencies (American Federation of Teachers, National Council on Measurement in Education and National Education Association, 1990). A parallel conceptualization is offered by Stiggins (1995), who asserts that "Assessment literate educators know the difference between sound and unsound assessment. They are not intimidated by the sometimes mysterious and always daunting technical world of assessment" (p. 240). According to Stiggins (1995), assessment-literate educators have a deep understanding of assessment. They know what they are assessing, why they are doing it, the best

ways to assess the skill or knowledge, how to create quality examples of student work, potential problems with the assessment, and how to fix those issues. They are also aware of the potential negative effects of poor, inaccurate assessment.

This way of defining allows being applicable in lots of conditions regardless of subject matter or other teaching conditions and facilitates the operationalize it in questionnaires or inventories. In contrast, researchers have recently criticized the conception of AL as a fixed set of knowledge and skills for several reasons (Willis *et al.*, 2013; Xu and Brown, 2016). To illustrate, according to Willis *et al.* (2013), AL represents a dynamic context-dependent social practice in which educators communicate and navigate classroom and cultural knowledge with each other and with students, as they engage in the implementation, evolution, and execution of assessment to support student learning objectives.

Correspondingly, Xu and Brown (2016) argue that being assessment-literate should involve developing some sophisticated abilities to navigate and engage with the complexities of assessment practices in diverse educational settings (Xu and Brown, 2016). Moreover, in their study, it is underlined that cognitive traits, beliefs and socio-cultural factors play a crucial role in developing AL that is a dynamic and context-dependent system. By moving beyond a narrow definition, the researchers aim to capture the multifaceted nature of AL, which includes aspects such as teachers' identity as assessors, critical examination of their practices, and participation in collaborative learning communities. Similar to this perspective, Pastore and Andrade (2019) consider the notion as a contextual concept that encompasses various interconnected components influenced by social, cultural, policy, professional, and experiential factors, beyond a list of specific knowledge or abilities.

In addition to this shift in the approach to AL, researchers restructure the components and scope of it as the research progress as well. For instance, although the initial definitions merely focus on summative assessment, further studies highlight the ability to develop and execute formative assessment methods (Popham, 2009; Yan and Pas-

tore, 2022; Xu and Brown, 2016). Additionally, Abell and Siegel (2011) emphasize that assessment literacy constitutes a fundamental component of teacher professionalism. This perspective has been further developed through subsequent research, with scholars such as DeLuca *et al.* (2010) explicitly linking AL to frameworks of professional growth for educators.

Collectively, recent research depicts AL as a dynamic, context-dependent, and socially situated construct, embedded within teachers' professional practices and shaped by their identities and perceptions of assessment (Meijer *et al.*, 2023). In contrast, initial approaches focused on listing specific abilities and knowledge to define the scope and align with assessment standards. In this study, none of the existing definitions were explicitly adopted to guide the research design. Instead, the dimensions comprising the assessment literacy framework were systematically extended from an existing conceptual framework.

1.3. The Multistage Testing

Several different assessment methods exist for gauging assessment literacy, including linear tests and adaptive tests. Both in the educational assessment field and other research areas, linear tests have wider representations due to their simpler development stages and administration conditions. Linear tests, such as paper-and-pencil tests (PPTs), employ a testing design where all test-takers encounter the same set of items, have the same test duration, and thus experience a uniform testing experience. In these test designs, the difficulty level of all items is predetermined, resulting in a test that may comprise a combination of easy, challenging, or varying difficulty level items that are consistent across all test-takers. Consequently, it can be challenging to differentiate the ability levels of the test-takers.

In contrast, the underlying principle behind adaptive testing is that a test provides more accurate and useful results when the difficulty of its items aligns well with the examinee's ability (van Der Linden and Glas, 2010). Thus, adaptive tests offer a

distinctive testing experience due to their unique design. In adaptive testing, rather than administering a uniform fixed-length test to all test-takers, the individual test-taker's proficiency estimate is iteratively updated after each new response or module, and the subsequent item or module is then selected to have optimal informational properties given the updated estimate (Van Der Linden and Glas, 2010). According to testing principles, adaptive tests can be divided into two categories as computerized adaptive tests (CATs) and multistage tests (MSTs).

Computerized adaptive testing represents an item-level adaptive approach, as it dynamically estimates the test taker's ability after each response and leverages this information to guide the selection of subsequent items from an item bank (Magis *et al.*, 2017). If the test taker answers a question correctly, the computer will select a more difficult item; conversely, if the examinee responds incorrectly, a less challenging question will be presented. This iterative process continues until a predetermined stopping criterion is met (Sari *et al.*, 2016). Therefore, compared to traditional linear tests, CATs are generally more efficient, offering lower measurement error across the full ability spectrum (Hendrickson 2007; Lord, 1980; Wainer 1990).

On the other hand, MST involves conducting tests in multiple stages, with the difficulty level and content of later stages adjusted based on performance in earlier stages instead of administering test items one at a time adaptively for each test taker through item-level adaptive testing in CAT (Yan *et al.*, 2016). The algorithm in MSTs administers pre-assembled groups of items (called modules) for each test taker, and the test is built up in stages. Thus, MST integrates principles from both CAT and linear testing, combining key features of these two assessment design approaches (Yan *et al.*, 2016). Since MST offers several benefits over traditional assessment methods and CATs by integrating content balancing, quality control, and adaptive design (Sari *et al.*, 2016), the popularity of MSTs has risen significantly in recent times.

In an MST, the stages consist of pre-assembled item sets, referred to as modules, designed to assess test takers at varying proficiency levels (Luecht and Sireci, 2011).

All test-takers initially complete a routing test at the first stage. Informed by their performance on this initial module, they are then directed to different test forms or modules tailored to their estimated ability level at the subsequent stage. The design parameters of multistage testing models, such as the number of stages, available modules per stage, and the size of the item sets within the modules may vary and typically pre-established (Yan *et al.*, 2014; Zenisky *et al.*, 2010). The administrative structure of MST, known as a panel, can also impact test forms. Modules within stages can be grouped based on administrative purposes (Yan *et al.*, 2016). Therefore, it is significant to determine design parameters according to specific test purposes and conditions.

A two-stage MST model featuring one panel is depicted in Figure 1.1. In this 1-2 design, the test starts from Stage 1 with the routing module which includes items in medium difficulty levels. Following that, the test takers in the second stage are either routed to the module including easy items or hard items according to their estimated ability levels. Thus, this design allows for two potential pathways. In standardized, operational tests multiple parallel panels allowing several pathways might be included to enhance test security and manage item and module exposure rates (Yan *et al.*, 2016).

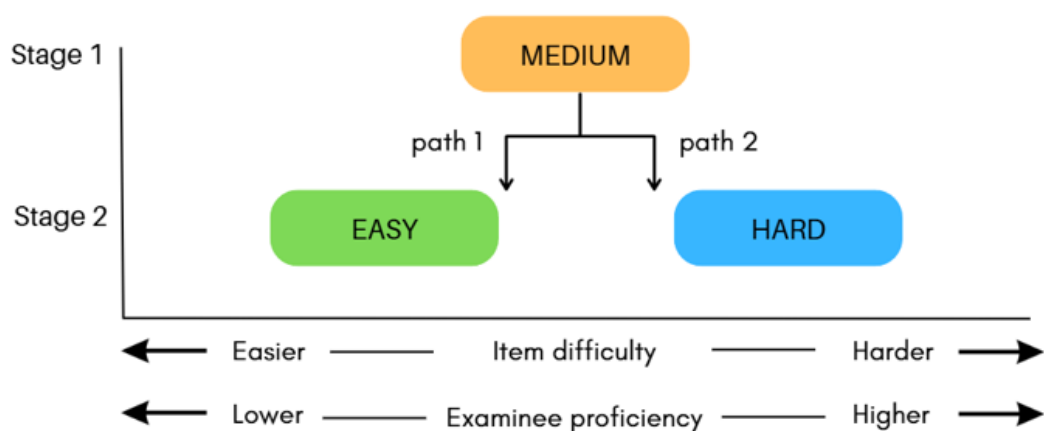


Figure 1.1. An example of 1-2 MST design.

In addition to fundamental design aspects such as test length, crucial factors to consider in test assembly for multistage testing also include adaptive routing algo-

gorithms, fair and accurate ability estimation methods, and optimizing test efficiency. MST models typically rely on the Item Response Theory (IRT) framework (Magis *et al.*, 2017). The IRT provides the foundation for item calibration, test design, module assembly, routing algorithms, and scoring procedures (Hambleton and Zenisky, 2013; Lord, 1971; Luecht, 1998).

The selection of the next module (routing) is determined by comparing the test taker's current estimated ability score to predetermined thresholds (or cut scores), and then applying specific rules to determine which module should be administered in the subsequent stage (Magis *et al.*, 2017). The widely used routing methods are number correct, Maximum Fisher Information and the Defined Interval Population. In addition to routing, the ability estimation method is another key component for MST design and the chosen routing method has an impact on ability estimation principles. To illustrate, if number-correct method was chosen for routing, ability estimation is required only to estimate final ability level at the end of the test while cut-off based, or information-based routing methods necessitate ability estimation to both calculate the final ability and to navigate the test takers (Sari *et al.*, 2016). Ability estimation methods can be grouped into two as Bayesian methods and non-Bayesian methods. The widely used methods can be summarized as shown in the following table:

Table 1.1. The summary of ability estimation methods.

Method	Type	Procedure	Advantage	Disadvantage
Maximum a Posteriori (MAP) (Lord, 1986)	Bayesian	Starts with an initial ability estimate and finds the ability level that maximizes	Easier mathematically than other methods	Can not estimate ability for perfect or zero scores, causing issues especially

Table 1.1 The summary of ability estimation methods. (cont.)

Method	Type	Procedure	Advantage	Disadvantage
		the likelihood function based on response patterns		with few items
Expected a Posteriori (EAP) (Hambleton and Swaminathan, 1985)	Bayesian	Uses the mean of the posterior distribution to estimate ability, based on a prior distribution	Non-iterative, easier to implement; does not require normal prior assumptions; lower bias and standard error compared to MLE and MAP	Accuracy suffers when the prior distribution is inappropriate
Maximum Likelihood Estimation (MLE) (Hambleton and Swaminathan, 1985)	Non-Bayesian	Multiplies a prior distribution with the likelihood function similarly to MAP but follows a non-Bayesian approach	Can provide ability estimates for perfect and zero scores	Can be heavily biased for small samples

All things considered, similar to CAT, multistage testing approaches aim to provide precise and accurate assessment of each examinee's proficiency across the full range of the ability distribution (Yan *et al.*, 2014). MST ensures precise assessment results by customizing content to individual abilities while enhancing the validity and reliability of assessments through quality control measures and IRT analysis (Sari *et al.*, 2016; Yan *et al.*, 2014). Furthermore, MST models also permit participants to revisit, review, and amend answers or leave them blank before advancing to the next stage (Yan *et al.*, 2014). These features have contributed to the widespread adoption of MST over CAT assessment practices (Magis *et al.*, 2017). Therefore, when creating an assessment literacy test, MST could be greatly advantageous in addressing a wide array knowledge and skills related to assessment literacy.

1.4. The Significance of the Study

The significance of the study lies in its dual contribution to the field of assessment literacy. Firstly, while existing frameworks (e.g., Pastore and Andrade, 2019; Xu and Brown, 2016) outline teachers' assessment knowledge and skills, the field lacks a comprehensive assessment framework. To address this critical gap, the current study offers an extension to a well-established conceptual framework, namely the teacher assessment literacy in practice (TALiP) framework (Xu and Brown, 2016) to reflect current educational demands, laying a practical foundation for developing instruments to measure assessment literacy. Second, this study underscores the need for novel measurement approaches that align with (1) recent advancements in educational assessment methodologies and (2) evolving teacher competency standards. Current research in the Turkey predominantly relies on measurement tools adapted from outdated standards, particularly the Ölçme Değerlendirme Okuryazarlığı Envanteri (hereafter referred to as the Turkish Assessment Literacy Inventory), developed by Bütünler *et al.* (2010), which was based on the Ministry of National Education's (the MoNE) teacher competencies released in 2005. Numerous studies (e.g., Azrak and Yalçınkaya, 2019; Erol and Tunagür, 2024; Koç and Bulut, 2020; Tünkler, 2019) utilized this measurement instrument despite significant transformations in assessment practices over the past

two decades. These transformations include the widespread integration of e-portfolios, digital testing platforms, computer-assisted item generation, and computerized adaptive testing systems. To respond to these critical issues, the current study pioneers the development of a MST format for measuring assessment literacy, drawing on the methodological advantages demonstrated by Yan *et al.* (2016). As the first MST designed specifically to measure assessment literacy among both preservice and in-service teachers in Turkey, this study makes significant contributions by: (1) advancing measurement precision through adaptive testing technology, and (2) providing empirical evidence for the utility of MST approaches in the specific domain of assessment literacy evaluation.

1.5. The Purposes of the Study

The current study has two primary purposes:

- to extend a well-established framework to a comprehensive assessment framework for assessment literacy by defining its core dimensions, ensuring alignment with current assessment practices
- to develop a multistage assessment literacy test for preservice and in-service teachers, based on current dimensions of assessment literacy.

1.6. Research Questions

This study investigated the following research questions:

- RQ1, What are the core dimensions of assessment literacy that align with current demands and practices of educational contexts?
- RQ2, To what extent does the multistage test developed to measure assessment literacy provide reliable and efficient measurement across preservice teachers' ability levels?

2. LITERATURE REVIEW

This chapter reviews existing frameworks for assessment literacy, research on measurement instruments for assessment literacy from around the world, and Turkish studies on assessment literacy.

2.1. Existing Frameworks for Assessment Literacy

The concept of assessment literacy (AL) has been explored in various studies (DeLuca *et al.*, 2016; Popham, 2009; Stiggins, 1995) however, it is only in recent years that researchers have started to systematically investigate its scope and propose models that outline its specific components. Existing literature features numerous studies dedicated to developing conceptual frameworks for AL from diverse perspectives. For instance, some researchers narrow their focus to specific subject areas, thereby constructing models with a more domain-specific approach.

In a recent study by Starck *et al.* (2018) a conceptual framework is proposed to help pre-service physical education teachers develop AL. The framework describes a four-step process for integrating assessment into teacher education. It offers suggestions to help physical education (PE) teacher educators progressively build pre-service teachers' assessment knowledge, with a focus on instruction and planning. The researchers do not ground their study in a specific definition of AL or established assessment standards; rather, they primarily consider Hay and Penny's (2009) conceptualization of assessment within the physical education field. According to Hay and Penney (2009), AL in physical education involves developing the knowledge and capacities to implement assessment and interpret its outcomes in a critically aware manner, thereby optimizing the value of assessment for all students.

Moreover, the occupational socialization theory "includes all the kinds of socialization that initially influence persons to enter the field of physical education and that

later are responsible for their perceptions and actions as teacher educators and teachers” (Lawson, 1986, p. 107) lays the foundations of this framework development study. Taking the domain-specific AL conceptualization for PE and the theory, a four-phased model is established by Starck *et al.* (2018), including four interrelated components of AL in PE context: assessment comprehension, assessment comprehension and application, assessment application and interpretation, and assessment interpretation and critical engagement with assessment.

The first phase, assessment comprehension, focuses on understanding assessment, including expectations and conditions for effective implementation. The goal is to grasp how assessment can support student learning, foster authentic learning experiences, and evaluate PE programs. In the second phase, researchers build pre-service teachers’ understanding of assessment concepts and guide them in applying assessment practices. This involves conducting actual assessments using either teacher-led or student-focused approaches. The goal is for them to practice implementing a range of assessment tasks in diverse settings and gather evidence of student learning. The emphasis is on expanding their toolkit of assessment tools to measure learning across different domains and using those tools for inclusive assessment. As pre-service teachers develop an understanding of the relationship between assessment, curriculum, and teaching and begin using effective classroom assessment strategies, they progress to the third phase. This phase focuses on applying assessment while interpreting assessment data. Pre-service teachers concentrate on making sense of and using the assessment information they have gathered, including navigating the social aspects of assessment while creating a plan to utilize the data from their school-based experiences. A key part of this phase is holding pre-service teachers accountable for the assessment knowledge gained in the initial stages and their ability to interpret assessment data related to instructional decision-making and student performance against established standards. The final phase integrates the previous stages and promotes critical examination of assessment practices. This phase encourages pre-service teachers to consider the power dynamics and social impacts of assessment, understand the limitations of data interpretations, and be cautious in how they utilize assessment results, acknowledging that all assess-

ments have inherent flaws and limitations (Hay and Penney, 2009). Consequently, the framework developed by Strack *et al.* (2018) presents a valuable examination of AL, effectively illustrating its multifaceted nature and the necessity of integrating technical skills with an understanding of sociocultural contexts. By utilizing Hay and Penney's (2009) AL framework, the study aligns with a growing body of research that demonstrates the framework's effectiveness in illuminating the strengths and limitations of teachers' assessment practices. For instance, another study focusing on primary PE teachers further validated Hay and Penney's framework (DinanThompson and Penney, 2015), revealing how teachers engaged with its four elements-comprehension, application, interpretation, and critical engagement. This study underscored the framework's utility as a tool for understanding and improving assessment practices, particularly in identifying gaps in teachers' knowledge and emphasizing the need for professional development to foster deeper critical engagement. Additionally, the framework provides a structured pathway for pre-service teachers to enhance their assessment practices, offering suggestions for building AL along with a learning focus for each component.

However, a critique arises regarding the framework's focused application specifically to physical education, rather than adopting a more holistic approach that encompasses broader educational contexts. By prioritizing context-specific challenges, the framework does not fully leverage opportunities for interdisciplinary learning and cross-contextual application. A holistic approach could encourage educators in various fields to engage with the principles of AL, thereby fostering richer dialogue across disciplines. In contrast, other studies of AL approach the construct from a broader perspective, examining AL in a general, holistic manner rather than focusing on domain-specific applications.

To exemplify, a framework of Teacher Assessment Literacy in Practice (TALiP) is proposed by Xu and Brown in 2016 to reconceptualize teacher AL by connecting educational assessment and teacher education research fields. Informed by sociocultural views of learning, the researchers define AL as a "dynamic, context-dependent social practice," emphasizing its situated and evolving nature within social and cultural

contexts. They view AL in its plural form, "assessment literacies," as contextualized and culturally responsive practices co-constructed by teachers and students in learning communities. This perspective encourages rethinking AL through diverse learning theories and assessment views, highlighting its interactive, relational, and participatory dimensions rather than treating it as a fixed set of skills.

The study's methodology involves the review of 100 studies on teacher AL from 1985 to 2015. Publications are identified and inductively analyzed, resulting in three main themes: (1) knowledge and skills within AL, (2) assessment education and its relationships with various mediating factors, and (3) contextual considerations of AL. Consequently, a new conceptual framework of teacher AL in practice (TALiP) is proposed integrating the various components and their interrelationships.

The TALiP framework comprises six components, arranged sequentially from the base to the top: the knowledge base, teacher conceptions of assessment, institutional and socio-cultural contexts, TALiP as the central concept, teacher learning, and teacher identity (re)construction as assessors. The foundational component, essential for effective teacher AL, is the "Knowledge base". It encompasses several critical elements that equip teachers with the necessary principles and competencies to conduct assessments competently. "Teacher conceptions of assessment" reflect their beliefs and perceptions about assessment, heavily influenced by prior experiences. "Institutional and socio-cultural contexts" define the broader environments in which assessments are conducted, including policies, cultural values, and social norms. "TALiP as the central concept" serves as the central idea around which the other components revolve, emphasizing the practical application of AL. "Teacher learning" highlights the continuous development of teachers through reflective practice and engagement in learning communities. "Teacher identity (re)construction as assessors" focuses on how teachers view themselves as assessors and how their experiences shape their assessment practices.

The TALiP framework redefines the traditional dichotomy of assessment "literacy" versus "illiteracy" by introducing a continuum that represents different degrees

of mastery (Xu and Brown, 2016). Additionally, it effectively highlights the intricate interplay between knowledge, teacher beliefs, contextual realities, and the cyclical development of assessment practices. However, while the framework is robust in its theoretical underpinnings, its practical applicability in diverse educational contexts remains underexplored.

In a similar vein to Xu and Brown (2016), Pastore and Andrade (2019) advocate for a holistic understanding of AL that includes teachers' beliefs, emotions, and contextual factors. However, while Xu and Brown (2016) propose a framework integrating sociocultural perspectives, Pastore and Andrade (2019) emphasize the socio-emotional dimension in teachers' assessment practice in their model. Their methodology involves a Delphi inquiry method which provides a structured process to gather expert opinions and reach a consensus on specific ideas (Creswell, 2014). By gathering international experts' feedback (35 first round and 27 second round) on the clarity, completeness, and usefulness of a new definition of teacher AL, a model is developed consists of three main components namely: Conceptual Dimension, Praxeological Dimension, and Socio-emotional Dimension.

The conceptual dimension encompasses an understanding of the fundamental principles and theories that undergird effective assessment practices. This includes knowledge of diverse assessment types, their intended purposes, and the capability to interpret assessment data in a meaningful manner. The praxeological dimension emphasizes the practical implementation of AL. It encompasses the competencies and approaches teachers require to enact effective assessment practices within their classrooms, including designing assessments, providing meaningful feedback, and utilizing assessment data to guide instructional decisions and learning. The socio-emotional dimension of AL examines the social and affective aspects of assessment, including the influence of assessment on students' learning experiences and emotional responses. This component underscores the significance of acknowledging the psychological dynamics inherent in assessment processes and developing the capacity to foster a supportive environment for students throughout these experiences.

Consequently, by articulating the interconnections between its three dimensions, the framework encourages teachers to reflect not only on their knowledge and practical skills, but also on their beliefs and values related to assessment (Pastore and Andrade, 2019). On the other hand, due to the Delphi method's focus on achieving consensus minority or dissenting viewpoints may be overlooked during the study.

Girgla *et al.*, (2021) also base their framework development study on a thorough review of existing literature. The unique contribution of this study on the field is that the study is not narrowed to a specific subject matter or grade level but rather structured the methodology to be able to support International Baccalaureate (IB) programs and to inform professional learning resources. As a result of the literature review, a scan of IB documents, and consultations with IB staff, the seven elements identify in the framework: assessment knowledge and skills; formative assessment; assessment identity; professional development; school environment; engagement of learners; and integrating digital assessment.

While Pastore and Andrade (2019), as well as Xu and Brown (2016), emphasize a multifaceted and dynamic conceptualization of AL, Girgla *et al.* (2021) build on this by integrating AL with design competence, particularly within the context of IB programs. Similar to Pastore and Andrade's three-dimensional model, Girgla *et al.*'s (2021) framework recognizes the interplay between conceptual knowledge, practical application, and the socio-emotional aspects of assessment. Likewise, their model aligns with the TALiP framework in acknowledging the evolving nature of AL within institutional and cultural contexts. However, a key distinction in Girgla *et al.*'s approach is its explicit focus on digital assessment and professional development structures tailored to IB educators. By incorporating elements such as formative assessment strategies, learner engagement, and school-wide assessment practices, their framework provides a structured yet adaptable model for enhancing AL within a globally recognized curriculum. This expanded perspective underscores the growing need for educators to navigate both traditional and digital assessment environments effectively. Overall, the study's integrative framework, emphasis on digital assessment contexts, and adoption

of a global outlook on AL constitute substantial advancements within the field of education, especially in the context of IB programs.

Building on the general frameworks for AL discussed earlier, some studies have sought to tailor these models to specific educational systems. For instance, the framework developed by Alzhrany (2022) is not only drawn on existing literature but also aligned explicitly with national teaching standards. A multifaceted approach is used including comprehensive literature review and an analysis of the National Professional Standards for Teachers in Saudi Arabia. Additionally, the inclusion of focus group discussions with local teachers to identify the essential assessment skills and knowledge required to meet the National Professional Standards for Teachers standards enhances the depth and nuance of the framework development study.

The insights from these discussions inform the creation of detailed “elaborations” that outline the specific competencies needed. Following this, the study designs a Likert scale survey to assess the perceived importance of these elaborations, and the feedback gathered is used to refine the framework further. Finally, a modified Teacher AL Questionnaire is developed, anchored in the finalized framework, to ensure alignment with the National Professional Standards for Teachers standards. This context-sensitive design ensures the articulated knowledge and skills are applicable and practical for the teachers it serves.

The frameworks discussed thus far have predominantly concentrated on the AL necessary for evaluating academic knowledge. However, the evolving educational landscape necessitates a more comprehensive approach. In this context, Chan and Luk (2022) have developed a framework that broadens the concept of AL to holistic competencies, thereby offering a more inclusive perspective on how educators can assess a diverse range of student abilities, including collaboration and creativity.

The researchers initially established a conceptual framework by identifying four central dimensions of teacher AL through a review of existing literature: knowledge,

attitude, practice, and socio-emotional management. They subsequently collected data through interviews and a focus group with university academics in Hong Kong, analyzing the responses thematically. This analytical process involves categorizing the data according to the identified dimensions. Ultimately, a four-dimensional model of teacher AL in holistic competency was established, comprising the dimensions of knowledge, attitude, practice, and socio-emotional management.

The knowledge dimension encompasses teachers' understanding of what to assess (target), why to assess (purpose), and the various assessment methods available (how). The attitude dimension includes teachers' beliefs, conceptions, and motivations regarding assessment, emphasizing the importance of an open-minded disposition and a willingness to question pre-existing beliefs. The practice dimension consists of teachers' proficiency in implementing effective assessment strategies, which entails the ability to define learning outcomes, select appropriate assessment tools, provide meaningful feedback, and assist students in understanding assessments. The socio-emotional management dimension focuses on teachers' capacity to address ethical issues in assessment and manage students' emotional responses to assessment processes.

Thus, it can be asserted that the general structure and primary components of Chan and Luk's (2022) model are consistent with previous frameworks. However, by expanding the scope of AL to include the assessment of holistic competencies rather than solely concentrating on academic knowledge, the framework's subdimensions become more specifically aligned with these holistic competencies.

2.2. Review of Global Research on Measurement Instruments for AL

The measurement of teacher AL has evolved significantly since the pioneering work of Plake *et al.* (1993), who developed the Teacher Assessment Literacy Questionnaire (TALQ). As one of the first standardized instruments to assess teachers' knowledge and skills in educational assessment, the TALQ focused on measuring in-service teachers' competency across the seven standards articulated in the 1990 Standards.

With this aim, an item pool including 35 multiple-choice items with five items corresponding to each standard was created, and the TALQ was administered to 555 in-service teachers and 268 administrators across 45 United States. The internal consistency reliability of the results, as measured by the Kuder-Richardson 20 (KR-20) coefficient, was 0.54 (Plake *et al.*, 1993). On average, the participating teachers answered slightly more than 23 out of the 35 test items correctly. Consequently, the TALQ has become a foundational instrument, influencing the scope and the design of subsequent instruments in the field.

In later years, the TALQ was renamed AL Inventory (ALI) by Campbell, Murphy and Holt (2002) and administered to 220 pre-service teachers. The findings indicated an internal consistency reliability estimate of 0.74 for the instrument. Additionally, the pre-service teachers in this study scored an average of 21, slightly lower than the average score from a previous study of in-service teachers (Campbell, Murphy, and Holt, 2002). Building on these two seminal works, Mertler (2003) developed a slightly modified version of the TALQ, labeled the Classroom Assessment Inventory (CALI), for investigating both pre-service and in-service teachers' AL. The CALI comprised the same 35 content-focused items, with only minor rewording (Mertler, 2003). The internal consistency reliability estimate was 0.57 for a sample of 197 in-service teachers, who scored an average of 22. Conversely, a sample of 220 pre-service teachers had a higher internal consistency reliability estimate of 0.74 but a lower average score of 19.

Therefore, the studies by Campbell *et al.* (2002) and Mertler (2003) replicated the work of Plake *et al.* (1993) by employing the same original measurement instrument and obtaining comparable findings despite differences in the samples. Following that Mertler and Campbell (2005) conducted research to develop an updated, context-specific version of the ALI. Although the researchers maintained the original allocation of 5 questions per standard, they distributed the 35 items across five different classroom scenarios that depict teachers who were facing various assessment-related decisions. This new instrument was called the Assessment Literacy Inventory (ALI) and administered on two separate occasions. The initial pilot testing included 152 preservice

teachers while the second phase consisted of 249 preservice teachers. The researchers concluded similar results in terms of psychometric evidence including reliability. On the other hand, it was underlined that formatting items as classroom scenarios might reduce cognitive overload with this approach when it compared to unrelated items (Mertler and Campbell, 2005).

These measurement instruments for assessing the AL of teachers and preservice teachers have served as foundations for subsequent researchers to design their methodologies. Several studies have systematically reviewed existing AL measures to evaluate their effectiveness and relevance (DeLuca *et al.*, 2016; Gotch and French, 2014; Pastore, 2023). For instance, Gotch and French (2014) conducted a comprehensive review of 36 AL instruments published between 1991 and 2012, categorizing them into objective tests, teacher self-reports, and rubrics. Their analysis focused on evaluating the psychometric properties of these measures against key claims, such as test content validity and internal consistency reliability. However, they concluded that the available psychometric evidence supporting these measures was limited, raising concerns about their adequacy for teacher evaluation.

Building on this work, DeLuca *et al.* (2016) expanded the scope of analysis by examining the alignment of AL measures with contemporary assessment standards across six English-speaking regions. Their review of eight measures published between 1993 and 2012 revealed that most instruments were based on outdated frameworks, such as the 1990 Standards, and failed to reflect modern practices like formative assessment and assessment for learning. While some measures demonstrated reasonable internal consistency, both studies converge in their conclusion that existing AL instruments suffer from significant limitations-whether in their psychometric properties or their alignment with current educational demands. A more recent review by Pastore (2023), synthesizing literature from 2013 to 2022, further underscores these limitations. Pastore highlights the continued reliance on outdated frameworks and the scarcity of innovative measurement instruments developed in the last decade. However, the review also identifies the Approaches to Classroom Assessment Inventory (ACAI), as a significant

advancement, aligning with contemporary educational standards and reflecting a more nuanced understanding of assessment practices.

In response to insights gained from earlier measurement instruments and reviews, DeLuca *et al.* (2016) developed a new assessment tool in 2016 called the Approaches to Classroom Assessment Inventory (ACAI). Unlike the previous instruments discussed, DeLuca *et al.* (2016)'s study centered on the Classroom Assessment Standards established by the Joint Committee on Standards for Education Evaluation (Klinger *et al.*, 2015) having significant differences with the 1990 Standards. To exemplify, the 1990 Standards are broader in scope and applicable to a wide variety of assessment contexts, including large-scale standardized tests and psychological evaluations. In contrast, the Classroom Assessment Standards for PreK-12 Teachers are specifically tailored for classroom teachers and aim to enhance assessment practices in order to improve student learning by emphasizing the importance of formative assessment, feedback, and student participation. Therefore, focusing on these standards might be considered as a distinctive aspect of DeLuca *et al.*'s work (2016).

In addition to the alignment of updated standards, the ACAI also features a distinct structure compared to the initial measurement instruments. Rather than creating one item for each standard as in the initial work, the structure varies across the three parts of the instrument: Approaches to Classroom Assessment, Confidence in Classroom Assessment, Assessment Professional Learning Priorities and Preferences. Part 1 includes five scenarios, with four items developed for each scenario corresponding to the four AL themes. In this part, teachers are asked to choose from three viable response options reflecting different classroom assessment approaches.

Part 2 uses a 5-point Likert scale for teachers to self-assess their confidence levels. Part 3 consists of two sections: one that uses a Likert-type scale from 1 to 5 for prioritizing learning interests, and another that uses a similar Likert scale from 1 to 5 for preferred professional learning modes. The ACAI was administered to a diverse sample of 404 Canadian teachers, including 63% in-service and 37% pre-service teachers. The

results suggested that DeLuca *et al.* (2016) provided a reliable and valid tool to capture teachers' diverse approaches to assessment.

Despite this progress, Pastore (2023) emphasizes the pressing need for new AL measures that are both psychometrically robust and aligned with the evolving complexities of AL in teaching practices. Collectively, these findings highlight the critical gap in the field and the urgency for developing instruments that better capture the dynamic nature of modern AL.

2.3. Review of Research on Measurement Instruments for AL in Turkey

While the global research on AL dates to 1990s, the first research that aims to examine assessment related knowledge and abilities in the context of AL in Turkey is an adaptation study conducted in 2010 based on the available literature. Bütüner *et al.* (2010) adapted the revised ALL, which was developed by Mertler and Campbell (2005), with the aim of developing an instrument in Turkish for measuring AL of preservice teachers. To make sure that the instrument is in line with Turkish educational assessment practices, Bütüner *et al.* (2010) examined the competency areas in the original instrument and eliminated the seventh competency including the use of assessment as an instructional intervention. By this competency area, Mertler and Campbell (2005) highlighted that need for “being skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information” (pg.21) and Bütüner *et al.* excluded the 5 items measuring this competency as this competency were not included in the MoNE’s 2005 teacher competencies. Thus, the researchers created the Turkish Assessment Literacy Inventory that is consisted of 30 scenario-based items which were originally developed based on the 1990s Standards.

The initial administration of the Turkish Assessment Literacy Inventory with 80 preservice teachers produced a KR-20 coefficient of 0.72, with mean item difficulty (0.59) and discrimination (0.38). Following revisions based on item analysis, the second administration with 180 preservice teachers demonstrated stronger psychometric prop-

erties, with $KR-20 = 0.86$, mean item difficulty = 0.64, and mean item discrimination = 0.49.

Bütüner *et al.*'s (2010) study can be viewed as the study paving the way for further research, as a significant amount of research on AL in Turkey utilized the Turkish Assessment Literacy Inventory based on the reviewed literature. For instance, after a relatively short period of time, Gül (2011) used the inventory with the purpose of investigating AL and attitudes towards assessment of preservice teachers from science education, social science education, Turkish education, and primary mathematics education departments. After implementing the instrument to 180 preservice teachers, it was concluded that the reliability of the results is acceptable ($KR-20 = 0.72$), while the participants demonstrated notably low competence based on the inventory.

In the following years, other studies using the Turkish Assessment Literacy Inventory consistently suggested a deficiency in assessment literacy among preservice teachers from various fields. Although the methodologies of these studies differ according to their research purposes, they indicate similar results regarding assessment literacy levels. For example, despite a large sample size ($N = 323$) compared to Gül (2011)'s study and a more diverse participant pool, the research conducted by Karaman and Şahin (2014) indicated low levels of assessment literacy among fourth-grade preservice teachers. Similarly, another study utilized the Turkish Assessment Literacy Inventory but was limited to a sample of preservice science teachers (Koç and Bulut, 2020). This study revealed that only 39.6% of the questions were answered correctly. Consistent with this, the findings of Azrak and Yalçınkaya (2019)'s study showed that preservice social science teachers answered only 33% of the questions in the inventory.

In addition to examining the current AL levels of preservice teachers, the researchers utilized the Turkish Assessment Literacy Inventory for additional research objectives. To illustrate, Tünkler (2019) explored the relationship between preservice social science teachers' AL levels and various other variables. Although the study did not find a statistically significant relationship between perceptions of assessment

self-efficacy and AL, the researchers did provide further evidence of gender-based differences in assessment literacy, consistent with findings from Karaman and Şahin (2014). Additionally, the study revealed no statistically significant differences in assessment literacy levels based on academic achievement. In a similar vein, another study found no statistically significant difference between AL and GPA (Koç and Bulut, 2020).

Unlike most existing studies mentioned earlier, Ergül and Çetin (2021) focused on in-service teachers. The researchers administered the inventory to 189 secondary school teachers from various branches and found that only 43% of the inventory questions were answered correctly. Notably, the results also indicated that the AL levels of participating in-service teachers significantly differed based on professional seniority and branch. Furthermore, using the Turkish Assessment Literacy Inventory, the researchers concluded that participants with at least 10 years of experience have lower AL levels than those with less than 10 years of experience. Although there is no other research in the Turkish context to compare these unexpected results, this finding contradicts previous research in a global context (Erdost, 2018; Plake *et al.*, 1993).

In contrast to the predominant use of the Turkish Assessment Literacy Inventory, there is another instrument development study conducted in Turkey. In the study by Ogan-Bekiroglu and Suzuk (2014), the primary research is stated as addressing pre-service physics teachers' AL and its implementation into practice. With this aim, the first author developed a novel measure by adapting some items from the questionnaire developed by McMillan (2001). The instrument diverges from the Turkish Assessment Literacy Inventory in both structural composition and foundational premise, and it was structured in two parts. The first part comprised 34 items that assessed preservice teachers' knowledge across three domains: assessment levels (e.g., application and reasoning), assessment methods (e.g., concept maps, portfolios), and grading standards (e.g., student effort, progress).

The second part consisted of 10 items that evaluated the teaching approaches of preservice educators, including lectures and group-based activities. According to

the results, pre-service physics teachers' AL was quite high, and Cronbach's alpha reliability value for the instrument was found to be 0.85 despite the small sample size (N=24).

The body of literature additionally contains a few studies on AL following a qualitative research paradigm (Demirdöğen and Korkut, 2021; Yaşar, 2020). For example, a recent multiple case study by Demirdöğen and Korkut (2021) investigated science teachers' AL across different majors (education and science). Based on content analysis deriving from several data collection tools, including interviews, observations, and documents, the study revealed that science teachers differ with respect to several dimensions of AL, such as views about learning, assessment purposes, assessment strategies, what to assess, and assessment interpretation. Another case study (Yaşar, 2020), which consisted of chemistry teachers, indicated that chemistry teachers lacked sufficient understanding and knowledge of diagnostic and formative assessment. Rather than providing quantitative evidence, Yaşar (2020) relied on 12 semi-structured interviews to assess teachers' formative assessment knowledge and perceptions, identifying significant deficiencies.

Overall, extensive investigation in Turkey into AL relies on using the Turkish Assessment Literacy Inventory as an instrument for various research questions. While most existing research adopts a quantitative paradigm and involves preservice teachers as the sample, a few studies focus on the AL of practicing teachers and employ qualitative methods. Although this body of literature provides significant findings and implications for the concept of AL, none of the studies critically examine the scope of the construct or frame their research around contemporary issues in assessment. For instance, in a study conducted in 2024 (Erol and Tuangür, 2024), data were gathered using the Turkish Assessment Literacy Inventory without any modification or adaptation, despite the instrument not addressing current needs in the educational assessment field, such as technology-based assessment. Therefore, it is reasonable to argue that further research is needed on the AL of both preservice and in-service teachers in Turkey.

3. METHODOLOGY

This chapter presents an in-depth examination of sampling method, participant characteristics, research procedures, and data analysis methods employed in the current study.

3.1. Participants

The participants of the study are experts for framework study and preservice teachers for test development phase. The participants in the assessment framework study included experts who volunteered to contribute to the study. The procedure for choosing experts involved a convenience sampling technique that entails collecting data from an easily accessible and available group of people (Creswell, 2014). This method was used for its efficient use in gathering different perspectives and insights quickly and efficiently. A total of 19 experts from diverse stakeholder groups, including academicians, assessment specialists, and instructors across various educational institutions, were engaged in this study. The demographic characteristics of the subject matter experts were presented below.

Table 3.1. Demographic characteristics of experts.

	N	%
<i>Gender</i>		
Female	13	68.42
Male	6	31.58
<i>Role</i>		
Associate Professor	8	42.11
Assistant Professor	4	21.05
Assessment Specialist	4	21.05
Professor	1	5.26
Educational administrator	1	5.26

Table 3.1 Demographic characteristics of experts. (cont.)

Teacher	1	5.26
<i>Institution</i>		
Ankara University	3	15.79
Hacettepe University	2	10.53
Terakki Foundation Schools	2	10.53
Balıkesir University	1	5.26
Bartın University	1	5.26
Boğaziçi University	1	5.26
Bilfen Schools	1	5.26
Erkan Ulu Schools	1	5.26
Gazi University	1	5.26
Anadolu University	1	5.26
Hakkâri University	1	5.26
İhsan Doğramacı Bilkent University	1	5.26
İstanbul Medipol University	1	5.26
The Teachers Academy Foundation (ÖRAV)	1	5.26
Ağrı İbrahim Çeçen University	1	5.26

The participants of the test development study consisted of undergraduate students enrolled in various programs at the education faculty, such as mathematics education, science education, and primary education. The selection process for participants involved using a non-probability sampling technique called purposive sampling. Purposive sampling involves selecting participants who meet specific criteria that are relevant to the research objectives (Creswell, 2014). The participants who have taken at least one assessment course were involved in the current study to ensure that they have relevant knowledge and experience in assessment principles and practices in education. The pilot implementation of the test in the paper-and-pencil format involved a sample of 1223 preservice teachers from 11 bachelor's degree programs across 11 universities. As shown in Table 3.2, the majority of the participants were female preservice teachers (72.2%) and from guidance and psychological counseling program (27.96%).

The participants for the MST consisted of 106 preservice teachers, drawn from 7 bachelor's degree programs across 3 universities. The demographic characteristics of these participants are outlined in Table 3.3.

Table 3.2. Demographic characteristics of participants for pilot.

	N	%
<i>Gender</i>		
Female	883	72.2
Male	328	26.82
Prefer not to say	12	0.98
<i>Program</i>		
Guidance and Psychological Counseling	342	27.96
Primary Mathematics Education	256	20.93
Primary Education	198	16.19
Turkish Education	129	10.55
Science Education	117	9.57
Social Sciences Education	112	9.16
Computer Education and Educational Technology	23	1.88
Turkish Language and Literature Education	16	1.31
Physics Education	13	1.06
Mathematics Education	10	0.82
Chemistry Education	7	0.57
<i>University</i>		
Balıkesir University	203	16.6
İstanbul University	194	15.86
Eskişehir Anadolu University	164	13.41
Yıldız Technical University	155	12.67
Boğaziçi University	114	9.32
Pamukkale University	103	8.42
Medeniyet University	98	8.01
Ankara University	91	7.44
Biruni University	51	4.17

Table 3.2 Demographic characteristics of participants for pilot. (cont.)

Başkent University	36	2.94
Middle Eastern Technical University	14	1.14
Total	1223	

Table 3.3. Demographic characteristics of participants for MST.

	N	%
<i>Gender</i>		
Female	83	78.30
Male	23	21.70
<i>Program</i>		
Science Education	40	37.74
Primary Mathematics Education	31	29.25
Turkish Language and Literature Education	17	16.04
Art Education	8	7.55
Mathematics Education	8	7.55
Philosophy Education	2	1.89
<i>University</i>		
Yıldız Technical University	53	50
Mimar Sinan Güzel Sanatlar University	34	32.08
Boğaziçi University	19	17.92
Total	106	

3.2. Procedure

3.2.1. The Development of an Assessment Framework for Assessment Literacy

The initial phase of the study involved the extension of a conceptual framework for AL to an assessment framework, drawing on a comprehensive review of existing frameworks in the literature and insights from expert interviews. The TALiP framework (Xu

and Brown, 2016) encompasses the core principles and components of AL based on an extensive review of numerous studies. Thus, the current study offered an extension of the TALiP by incorporating assessment framework approach and proposing new dimensions. Additionally, the teacher competencies published in 2017 by the MoNE and several assessment textbooks were carefully considered during this process. Building upon the knowledge base of the TALiP framework, as well as examining other existing frameworks and the teacher competencies, a draft of the assessment framework was created. The dimensions and scope of the TALiP were not merely utilized but rather modified to suit the specific needs of this study. Consequently, a draft of the framework was developed to discuss with experts. The draft included eight dimensions, each accompanied by its corresponding subdimensions.

Following the development of the draft, experts were reached through emails. Experts who had consented to take part in the study were sent to an informative document including the purpose of the study, the draft framework and interview questions to increase the productivity of meetings. Semi-structured interviews were conducted individually, and experts were asked a couple of questions as follows:

- What are your thoughts on the assessment literacy dimensions described? Do you believe the listed dimensions effectively capture the scope of assessment literacy?
- How do you evaluate the subdimensions provided for each dimension? Do you feel they adequately cover the dimension they represent?
- Based on your expertise, is there any dimension or subdimension you would recommend relocating or introducing a new one?

As the interviews progressed, conflicting opinions were specifically discussed, and additional questions were added to the informative document. Each expert interview was recorded and transcribed. The suggestions and questions derived from the expert opinions were analyzed and synthesized with the findings of the literature review to finalize the assessment framework. This phase of the study can be summarized as in the following figure:

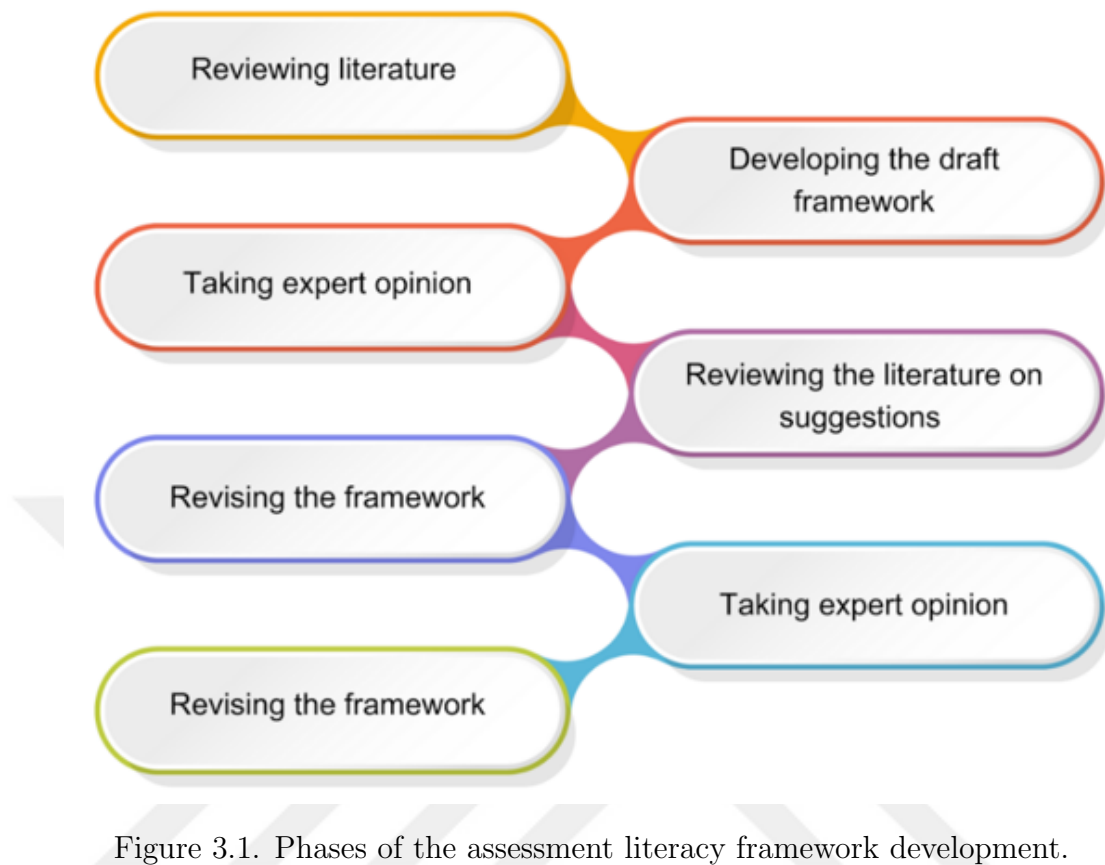


Figure 3.1. Phases of the assessment literacy framework development.

3.2.2. The Test Development for Measuring Assessment Literacy

The preliminary phase in creating the MST for assessment literacy entailed developing and administering a paper-and-pencil test during the pilot testing stage. This enabled the acquisition of item parameters and the assessment of item quality before designing the multistage test.

3.2.2.1. Pilot testing. Crocker and Algina (2006) define the primary goal of test development as a subjective-centered approach that aims to locate individuals on a quantitative continuum with respect to a particular psychological construct. Since the purpose of the test development in the present research aligns with Crocker and Algina's approach, their systematic 10-phase framework for test construction was followed.

- *Identifying the purpose of test score use*; The primary purposes and priorities for using the test scores are crucial factors that can significantly impact the likelihood of the final version being beneficial for its intended use (Crocker and Algina, 2006). The fundamental goal of the current study is to develop an MST for measuring AL. Although the intended use of the test may change after the final form is created, such as becoming a summative assessment tool for undergraduate assessment courses or a diagnostic test in measurement and evaluation centers where assessment specialists are employed, the test is ultimately designed to evaluate the test takers' AL competencies.
- *Identifying behaviors to represent the construct*; To avoid a biased and personalized understanding of the construct, it is advised for test developers or researchers to employ diverse methods to identify behaviors intended to reflect the construct of interest (Crocker and Algina, 2006). Since the assessment framework developed in the current study laid the foundation for the test, all items were generated based on measurable definitions of behaviors related to the dimensions.
- *Preparing test specifications*; After determining the objectives of the test, creating a definition, and outlining behaviors, it is important to establish a plan for evaluating the appropriate level of significance for each of these components in the test (Crocker and Algina, 2006). Therefore, in this stage, a table of specification was developed.
- *Construct an initial pool of items*; The first step in developing the item pool was determining the structure and format of the items. According to Popham (1981), item formats can be divided into two main categories: those that require the respondent to create a response, such as essay questions, and those that require the respondent to choose a response from several options, such as True/False items. Downing and Haladyna (2006) referred to the latter as selected-response items, which they considered highly beneficial and appropriate for assessments intended to make inferences across a wide range of knowledge, skills, or cognitive abilities. Therefore, 4-choice multiple-choice items, a form of selected-response items, were used in this study to assess pre-service teachers' AL competencies. This widely used item format is chosen for multiple-choice assessments due to

its versatility, effectiveness in evaluating various levels of knowledge and skills, suitability for both traditional and computerized testing methods, and ease of production, validation, storage, and recycling during the test creation process (Downing and Haladyna, 2006).

Furthermore, the item pool was crafted to assess a range of assessment literacy competencies, not merely basic knowledge, by encompassing items that evaluate both the knowing and applying cognitive domains. In line with the assessment framework for mathematics developed by the Trends in International Mathematics and Science Study (TIMSS), the knowing cognitive domain included items that assess abilities such as recalling facts, identifying characteristics, ordering the steps of a process, and defining concepts (Mullis *et al.*, 2021). Conversely, the applying cognitive domain encompassed items that require the ability to apply knowledge and conceptual understanding in a range of contexts (Mullis *et al.*, 2021). For this cognitive domain, items necessitated the abilities of computing, implementing, and transferring knowledge into practical applications. On the other hand, the third cognitive domain in the TIMSS framework could not be integrated because it is challenging to measure reasoning abilities, such as generalizing and justifying, through structured multiple-choice items. Additionally, at this phase, the guideline for developing selected-response test items proposed by Haladyna and Rodriguez (2013) was followed to increase the validity of test score interpretations.

- *Have items reviewed (and revise as necessary)*; Once the item pool has been established, a group of experts were asked to provide feedback on the items by reviewing elements such as precision, suitability or alignment with test specifications, technical flaws in item construction, grammar, potential for bias or offensiveness, and readability level.
- *Hold preliminary item tryouts (and revise as necessary)*; Prior to finalizing the items for a field test, it is recommended to pilot them with a small group of examinees. Twelve undergraduate teachers, who volunteered to contribute to the study, were asked to solve one booklet containing 20 items and provide feedback on their testing experience. Additionally, a debriefing session was held to en-

courage participants to share their thoughts on each question and propose ideas for potential enhancements. During this phase, 22 items in the item pool were modified based on students' suggestions and corrections, including revisions to the arrangement of visuals, wording, and strengthening of distractors.

- *Field-test the items on a large sample representative of the examinee population for whom the test is intended*; When the item pool consists of a substantial number of items, it is recommended to develop booklets utilizing Multiple Matrix Sampling (Gonzalez and Rutkowski, 2010). Therefore, a Multiple Matrix Sampling (MMS) design was implemented at this stage, enabling the regulation of item exposure, reducing the burden on students, and improving the efficiency of the assessment process (Aybek *et al.*, 2024; Shoemaker, 1973). In accordance with the MMS framework, seven booklets were created, as illustrated in Table 3.4.

Table 3.4. MMS design structure.

Booklet	Blocks		
1	A	C1	C2
2	B	C2	C3
3	C	C3	C4
4	D	C4	C5
5	E	C5	C6
6	F	C6	C7
7	G	C7	C1

As shown in Table 3.4, this MMS design links booklet 1 to booklet 2 through C2, and booklet 2 to booklet 3 through C3, and so on. To create an item pool of 105 items, the non-rotating section included 10 items, while the rotating anchors comprised a total of 10 items, such as C1 and C2, ensuring calibration for all 105 items. Consequently, each participant was required to complete one booklet containing 20 items within 30 minutes.

- *Determine statistical properties of item scores and, when appropriate, eliminate items that do not meet pre-established criteria*; Analyzing the statistical charac-

teristics of item scores is crucial for ensuring the validity, reliability, fairness, and efficiency of a test. These processes enhance the overall quality of the assessment tool and improve the accuracy of inferences drawn from test scores (Crocker and Algina, 2006). During this phase, item analysis based on IRT was conducted to select items that have the greatest impact on the test score and eliminate those that fail to meet standards for being classified as satisfactory or acceptable before finalizing the test. Further details are provided in the Data Analysis section.

- *Design and conduct reliability and validity studies for the final form of the test;* After administering the test, it is crucial to perform reliability and validity analyses on the test scores. Reliability analysis yields evidence regarding the consistency of interpretations across diverse contexts, whereas validity analysis provides evidence that the test truly measures what it aims to assess conceptually (DeVellis, 2017). The Data Analysis section provides details on the analyses conducted at this stage.
- *Developing guidelines for administration, scoring, and interpretation of test scores;* Once a test is ready for use following the required analysis, it is recommended to develop guidelines to facilitate its administration, scoring, and interpretation (Downing and Haladyna, 2006). However, within the current study, this phase could not be implemented due to time constraints.

3.2.2.2. The Development of Multistage Test. Following the pilot testing data analysis, the MST test was designed. The item analysis conducted using the IRT allowed for item selection based on pre-calibrated item parameters and item-fit indices. Each module was carefully designed to include one item representing each dimension of the assessment framework, thereby ensuring content balance across the modules.

The sample items presented in Figure 3.2, Figure 3.3, and Figure 3.4 were placed in the second stage of the MST. These three items aimed to assess the participants' competency in interpreting and communicating the data obtained from assessment results. Specifically, these items focused on the standard error of measurement, which should be interpreted. As can be seen from the sample items, the difficulty did not stem

from the need for competency in other areas or extraneous factors such as complexity of wording. Instead, the items necessitated more comprehensive knowledge regarding the same content. According to the IRT analysis conducted during the pilot phase, their difficulty parameters were calculated as -0.34, 0.79, and 1.26, respectively. For the third stage, another three items having varying difficulty levels were placed into easy, medium, and hard modules for another objective under the same competency. Rather than setting strict b parameter intervals for each module or stage, the obtained b parameters were carefully examined during item placement for each item objective. This test design approach ensured content balance.

Semih Öğretmen uyguladığı yazılı sınav puanlarına ait ölçmenin standart hatasını hesaplamak istiyor.

Bu hesaplamayı yapabilmek için sınavla ilgili hangi bilgilere ihtiyaç duyar?

- A) Standart sapma- Madde Güçlük İndeksi
- B) Aritmetik Ortalama- Madde Güçlük İndeksi
- C) Standart sapma- Güvenirlilik katsayısı
- D) Aritmetik Ortalama- Güvenirlilik katsayısı

Figure 3.2. Sample item 1.

Emre Öğretmen, İngilizce sınavından alınan puanlara ait aşağıdaki bilgileri hesaplamıştır:

Standart Sapma	4
Güvenirlilik Katsayısı	0,64

Buna göre İngilizce sınav puanlarına ait ölçmenin standart hatası kaçtır?

- A) 46
- B) 34
- C) 3,2
- D) 2,4

Figure 3.3. Sample item 2.

Nilay Türkçe sınavından 80 puan almıştır.

Bu sınava ait ölçmenin standart hatası 5 olduğuna göre, Nilayın gerçek puanı %95 olasılıkla hangi puan aralığındadır?

- A) 60 - 100
- B) 65 - 95
- C) 70 - 90
- D) 75 - 85

Figure 3.4. Sample item 3.

As a result, the MST was manually assembled to construct the optimal test structure. The 1-3-3 module structure was preferred because it enables two adaptive decisions, which enhances measurement precision while keeping the test length reasonable (Zheng, 2015). The test was designed to begin with stage 1, where all participants encountered a common set of moderately challenging items known as the routing test. Subsequent to stage 1, each participant's ability level is computed, and the most informative module for stage 2 is selected based on the results from stage 1. After stage 2, the process of estimating ability level is repeated, and a module for stage 3 is chosen for administration. The test concludes after stage 3, at which point a final ability estimate is determined. The MST structure is shown in Figure 3.5. Consequently, the MST employed a single panel structure, which included 7 modules and a total of 56 test items.

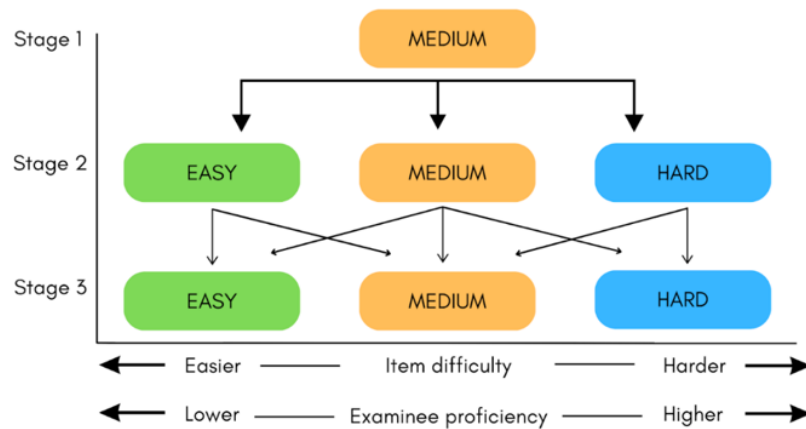


Figure 3.5. Three-stage 1-3-3 MST design.

In an IRT-based MST, it is generally preferable to use dynamic rules that include selecting the next module using an information-based criterion, such as one that maximizes Fisher information (Magis *et al.*, 2017; Weissman *et al.*, 2007). In IRT, information reflects an item’s ability to precisely discriminate around the examinee’s current ability estimate. Thus, selecting the module that maximizes information at the current ability estimate (θ) ensures higher measurement precision and efficient reduction of measurement error (Yan *et al.*, 2016). In the Rasch and two-parameter logistic models, items provide the most information when the difficulty parameter matches the examinee’s latent trait, which corresponds to a 50% probability of a correct response (Magis *et al.*, 2017). Static routing methods compromise the Approximate Maximum Information method (Luecht, 2000) mainly used in criterion-referenced test administration, the Defined Population Interval Method (DPI) generally used in norm-referenced test administration (Zenisky, 2004) and the number-correct method.

In the current study, the Defined Population Method was employed to guide the selection of modules for test takers. The main goal of this method is to direct examinees to modules to ensure that an equal or highly similar number or proportion of examinees follows each possible pathway (Sari *et al.*, 2016). In the 1-3-3 panel structure, the θ values corresponding to the 33rd and 67th percentiles of the cumulative θ distribution were calculated as -0.44 and 0.44, respectively. These values serve as cut points in the DPI method (Sari *et al.*, 2016). The estimated true scores are then calculated from these θ values. Finally, examinees are routed to alternative modules (easy, medium, hard) based on the number of correct responses they received in the current module.

For ability estimation, the expected a posteriori method (EAP) was favored due to its advantages, such as providing finite estimates and requiring less computational resources (Embretson and Reise, 2000). The MST was developed and delivered using the Concerto Platform, an open-source online adaptive testing platform created by the University of Cambridge Psychometrics Centre.

3.3. Data Analysis

3.3.1. Reliability

Reliability is a fundamental psychometric characteristic that captures the overall consistency and stability of the scores obtained from a test or measure (DeVellis, 2017). Reliability analysis under the classical true score model involves estimating the true score of test-takers and determining the extent to which the obtained test score represents the true score. Test results can be considered highly reliable if there is a strong correlation between respondents' observed scores and their actual level or amount of the attribute being measured. Additionally, the reliability coefficient represents the proportion of observed score variance attributable to true score variance (Crocker and Algina, 2006).

There are different ways to compute the reliability coefficient based on the decision intended to be made, such as the inter-rater, test-retest, or internal consistency method. The interrater approach involves two raters scoring the same test results to ensure consistency across raters, while the test-retest method includes administering the same test to a group on two different occasions. In the current study, the internal consistency of test results was examined to determine the degree of consistency of test takers' performance across all items within a test which are designed to measure the same underlying construct (Crocker and Algina, 2006).

Cronbach's alpha (α) is a commonly utilized measure of internal consistency, with values ranging from 0 to 1. Stronger internal consistency is suggested by higher Cronbach's alpha values. The alpha coefficient is computed based on the average inter-item correlation across all items in the test, and formulated as follows,

$$\alpha = \left[\frac{k}{(k-1)} \right] \times \left[1 - \left(\frac{\sum_{(i=1)}^k (V X_i)}{V(O)} \right) \right], \quad (3.1)$$

where k is the number of items, $V(X_i)$ is the variance of responses to item i and $V(O)$ is the variance of the observed sum (Cronbach, 1951).

A Cronbach’s alpha ranging from 0.70 to 0.80 is considered “acceptable”; between 0.80 and 0.90 is considered “good”; and above 0.90 is regarded as “excellent” (Cicchetti, 1994).

On the other hand, while Cronbach’s Alpha has historically been popular and extensively utilized, there is an increasing belief among psychometricians and researchers that other methods for reliability estimation have the potential to represent more precise measurements in practice (Dunn *et al.*, 2014, Hayes and Coutts, 2020, Padilla *et al.*, 2013). McDonald’s Omega (ω), introduced by McDonald in 1999, is widely preferred over Cronbach’s Alpha for various reasons. To exemplify, since Omega does not assume that all items in a test have equal factor loadings and measure the latent variable with the same degree of precision as Alpha, its reliability estimation is regarded as more accurate when the assumption is not met (Hayes and Coutts, 2020). Hence, apart from Alpha, McDonald’s Omega was computed to draw conclusions about the internal reliability of test results in the present study.

In addition to these two coefficients based on the Classical Test Theory (CTT) approach, the overall marginal reliability of the entire item bank was also evaluated (Cheng, Yuan and Liu, 2012). Theoretically, the marginal reliability is calculated via the following formula, where $E[SE(\theta)^2]$ is equal to the average squared standard error of ability level/theta estimates and σ_θ^2 is equal to the variance of ability level/theta estimates (Kim, 2012),

$$\text{Marginal Reliability} = 1 - \frac{E[SE(\theta)^2]}{\sigma_\theta^2}. \quad (3.2)$$

The study utilized the psych package (Revelle, 2023) to compute both Cronbach’s Alpha and McDonald’s Omega, and the mirt package (Chalmers, 2012) for marginal reliability in the R Software (R Core Team, 2023; version 4.2.3).

In contrast to linear tests, the reliability of scores from a MST varies across the test takers’ ability levels, as tailored item sets are administered to individual test-takers based on their estimated ability (van Der Linden *et al.*, 2014). Thus, IRT-based

reliability analysis, which acknowledge that measurement precision varies by ability level, is recommended to report as reliability evidence of adaptive test results (van Der Linden *et al.*, 2014). In the current study, the reliability of the MST was investigated by two IRT-based approaches: expected a posteriori (EAP) reliability and marginal reliability indices.

For the EAP reliability, the item parameters for each module were obtained from pilot testing phase and then used to compute the test information functions (TIFs) across a range of ability levels. EAP reliability was estimated that involved integrating over a standard normal distribution of ability. The posterior variance was approximated as the reciprocal of the test information function, and the EAP reliability was calculated as the ratio of the variance of expected ability estimates to the total variance (Kim, 2012). Additionally, marginal reliability indices were calculated separately for each stage and module by integrating the test information function over the normal ability distribution, providing insight into the expected precision of measurement at the population level. All analyses for reliability of the MST results were conducted using R Software (R Core Team, 2023; version 4.2.3).

3.3.2. Validity

Validation is defined as the process by which a test developer or test user gathers data to support the different types of inferences that are to be derived from test results (Messick, 1989). The type of validation study differs in terms of the desired inference (Crocker and Algina, 2006). According to the intended inference, there are essentially three types of validity: content validity, construct validity, and criterion-referenced validity.

For pilot testing, the content and construct validity of the MST were examined as part of this research. Content validity refers to the extent to which a specific set of items reflects a content domain (Devellis, 2017). Therefore, the purpose is to assess whether the items adequately represent a performance domain or construct of specific

interest (Crocker and Algina, 2006). In the current study, content validation evidence for pilot testing was collected through a table of specifications, expert opinions, and the index of item-objective congruence index.

The table of specifications outlined the distribution of items across different dimensions and cognitive levels for each test booklet. This ensured comprehensive coverage of the intended content and a balanced assessment of the different aspects of assessment literacy. Additionally, the index of item-objective congruence (IOC) described by Rovinelli and Hambleton (1977) was evaluated to collect supportive data for previous content-validation studies and quantitatively summarize the experts' opinions. The following congruence index formula enables assessment of how well an item aligns with the test's objectives, providing valuable information on item validity (Crocker and Algina, 2006),

$$IOC = \left(\frac{N}{2N - 2} \right) \times (\mu_k - \mu). \quad (3.3)$$

To be able to compute the index, experts were requested to match the item to each objective and assign a value of +1 if it measures the objective, 0 if uncertain, and -1 if it clearly does not match. The formula given above was used with N representing the total number of objectives, μ_k denoting the judges' average rating of item i on the k th objective, and μ representing the judges' average rating of item i on all objectives. In this case, the total number of objectives, which is N in the formula, represents the number of dimensions in the assessment framework.

For these calculations, an MMS design containing seven booklets was generated as shown in Table 3.5. This design for calculations allowed for decreasing the cognitive load of the experts' and gathering at least 2 expert's opinion for each item. Each expert rated one booklet that was sent as Google Form link. Furthermore, to ensure the reliability and validity of ratings, experts were sent to an informative video about the IOC index, how to rate the items and the framework including subdimensions. Consequently, each expert evaluated approximately 35 items in total. The number

of content experts was considered when determining a criterion for item acceptance (Rovinelli and Hambleton, 1977; Turner and Carlson, 2003). Items that demonstrate incompatibility at this point were carefully examined, modified, or eliminated from the test before MST design accordingly to ensure validity.

Table 3.5. MMS design for IOC calculations.

Expert	Booklet	Booklet	Total number of items
E1	B1	B2	34
E2	B1	B4	35
E3	B1	B5	35
E4	B3	B4	35
E5	B3	B5	35
E6	B3	B6	35
E7	B2	B5	36

Besides content validation, the construct validation study was conducted for the pilot test results. To assess construct validity, initially, an Exploratory Factor Analysis (EFA) was conducted to examine the underlying factor structure of each booklet's test results. EFA was performed by using IBM SPSS Statistics (Version 29). To conduct EFA, the Principal Axis Factoring extraction method and oblimin rotation was utilized. During the analysis, factor loadings that are less than 0.30 were suppressed. To ensure the feasibility of EFA, the Kaiser-Meyer-Olkin measure and the Bartlett's Test of Sphericity were examined for each booklet.

Another widespread way for providing validity evidence is conducting Confirmatory Factor Analysis (CFA) to examine how well the collected data fits the proposed model (Ullman, 2006). The five steps of the CFA were described as follows: model specification, model identification, model estimation, model testing, model modification (Ullman, 2006).

The first step of CFA involves reviewing the existing research literature to determine which model is most appropriate for examination and validation (Ullman, 2006).

In this study, the model was determined based on the assessment framework including eight dimensions which were derived from literature synthesis and expert meetings. In the second stage, the identification of a model occurs when the sample covariance matrix and the theoretical population covariance matrix yield a unique solution for each model parameter, and the number of estimated parameters exceeds the number of data points (Ullman, 2001; Ullman, 2006). The next step entails estimating the model parameters, particularly the factor loadings and error variances. There are several model estimation methods including Maximum Likelihood Estimation, Weighted Least Squares. After specifying, identifying, and estimating the model parameters, the next phase is to evaluate the model based on fit indices. The Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) are employed for evaluating model fit (Ullman, 2006). RMSEA values span from 0 to 1, where values below 0.06 are considered indicative of a good fit, while values exceeding 0.10 are interpreted as indicative of a poor fit (Ullman, 2001). Both CFI and TLI also range from 0 to 1, with values above 0.95 signifying a good fit (Hu and Bentler, 1999; Ullman, 2001). In the final step, if the model's fit indices are found to be unsatisfactory, it is best to modify the model to improve its fit based on sound theory and logical reasoning (Ullman, 2006).

3.3.3. Item Response Theory

Classical test theory (CTT) has long guided test development, but contemporary measurement increasingly favors Item Response Theory (IRT) for its enhanced psychometric properties (Embretson and Reise, 2000). CTT relies on total scores and assumes uniform accuracy across test-takers. On the other hand, IRT estimates trait scores and standard errors individually for each response pattern while accounting for item characteristics (Baker, 2001). Notably, CTT's parameters (e.g., difficulty, discrimination) are fixed but sample-dependent, whereas IRT models estimate separate ability and difficulty parameters for each item and test-taker, making them both adaptive and sample-independent (Baker, 2001, 2004; Hambleton and Swaminathan, 1985).

Therefore, the present study employed the IRT approach, taking advantage of its suitability for adaptive testing.

Before employing an IRT model, there are some theoretical assumptions that need to be checked whether they are violated or not. The IRT models rely on three fundamental assumptions: monotonicity, local independence, and unidimensionality.

The first assumption is referred to as the monotonicity assumption which implies that as the ability level increases, the probability of getting the item correct increases (Embretson and Reise, 2000). In the current study, the monotonicity assumption was checked via Item Characteristics Curves (ICCs), which are smooth S-shaped curves describe the relationship between the probability of correct response to an item and the ability scale (Baker and Kim, 2004). As illustrated in Figure 3.6, the probability that an examinee with a specific ability level will answer an item correctly is represented by $P(\theta)$. For a typical test item, this probability tends to be low for examinees with lower ability levels and high for those with higher ability levels.

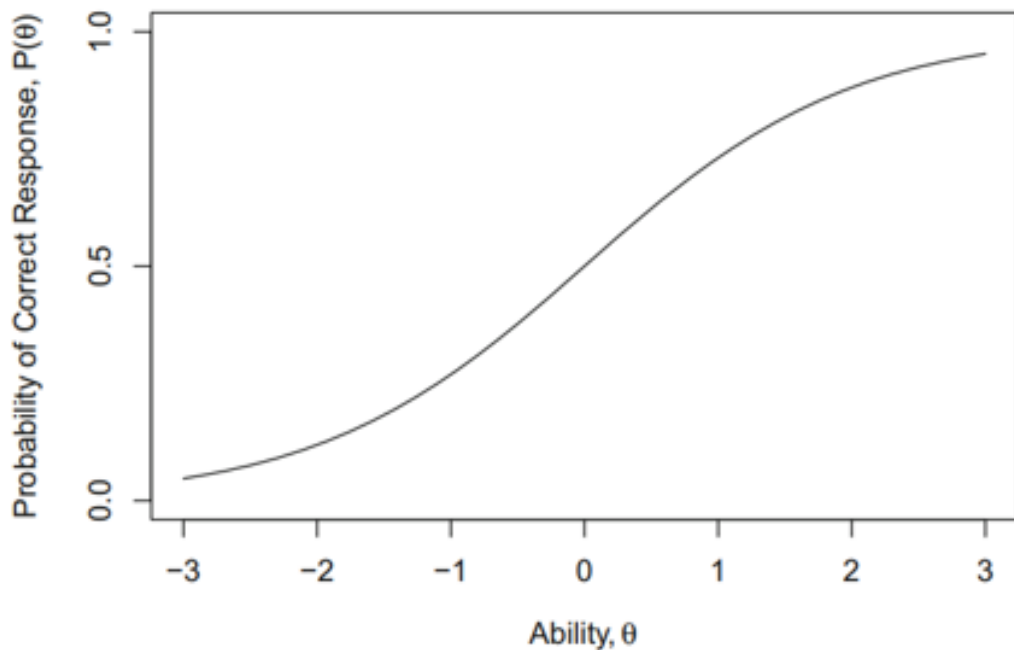


Figure 3.6. A typical item characteristic curve.

Furthermore, local independence is achieved when the probability of solving any item is independent of the outcome of any other item, while controlling for person parameters and item parameters (Embretson and Reise, 2000). To test local independence Yen's Q3 test was utilized, which computes the correlations between item residuals after fitting the IRT model (Yen, 1984). Both the ICCs per item and Yen's Q3 statistics were created by using R Software (R Core Team, 2023; version 4.2.3).

The third assumption of IRT models is unidimensionality, suggesting that only one latent trait is measured by the set of items in the test. There are several ways to check this assumption including determining the number of eigenvalues greater than 1, which is also known as the Kaiser rule, (Kaiser, 1960), parallel analysis (Horn, 1965; Green *et al.*, 2012), the scree test (Cattell, 1966), and the ratio of the first and second eigenvalues as an index of unidimensionality (Lord, 1980). To check this assumption, the Kaiser rule, parallel analysis (Horn, 1965; Green *et al.*, 2012), and the first-to-second eigen value ratio rule (Lord, 1980) was examined.

Under the IRT framework, there are various models based on their estimated parameters. The Rasch model, which is also known as the one-parameter logistic (1-PL) model, uses the following function to model the probability of success on an item with item difficulty and ability, where b is the difficulty parameter and θ is the ability level (Rasch, 1960),

$$P(\theta) = \frac{1}{1 + e^{-1(\theta-b)}}. \quad (3.4)$$

Thus, the discrimination parameter is fixed at a value of a (the item discrimination parameter) = 1.0 for all items under this model. The b parameter is defined as the ability level needed to have a 0.50 chance of correctly answering an item when assuming a guessing parameter of zero (Baker, 2001). A higher b -parameter value indicates a more challenging item, while a lower value signifies an easier item.

Unlike the 1-PL model, the two-parameter logistic (2-PL) model also includes the item discrimination parameter which is denoted by ‘a’. The parameter “a” indicates the degree to which a test item discriminates between examinees at different ability levels (Baker, 2001). In addition to the a and b parameters, the three-parameter logistic (3-PL) model also estimates the parameter c, representing the probability of getting the item correct by guessing alone. In the current study, the Rasch model was utilized instead of more complex models (e.g., 2PL or 3PL) due to the limited sample size, which ensured parameter estimation stability and helped avoid overfitting (Baker and Kim, 2004).

Examining item fit indices is essential to assess the appropriateness of each item for the model (Embretson and Reise, 2000). In this study, Orlando and Thissen’s Chi-Square Item Fit values, also known as the S-x2 statistic, were computed to determine whether the items fit with the model (Orlando and Thissen, 2000). This IRT-based statistic was favored since it has been shown to be more robust to small sample sizes or extreme item difficulties (Orlando and Thissen, 2000). The S-x2 statistic for each item in pilot testing was computed by using the ‘mirt’ R package (Chalmers, 2012) via Rstudio (Rstudio Team, 2020). Additionally, to assess item fit to generated IRT model, the ICCs were analyzed since systematic deviations in the ICCs might indicate potential misfit.

In the IRT, the information functions at both the item and test levels are essential as they offer valuable insights into measurement precision (Baker, 2001). At item level, the amount of information, based upon a single item, can be computed at any ability level and an item provides maximum measurement precision when the examinee’s ability level (θ) equals the item’s difficulty parameter (Baker, 2001). As ability deviates from this point, either above or below, the item’s information diminishes, becoming negligible at the extremes of the ability distribution (Baker and Kim, 2004).

At test level, the IRT enables to create the test information function (TIF) to evaluate the overall effectiveness of the test in measuring ability across a range of

ability scores (Baker and Kim, 2004). The TIF at a given ability level is simply the sum of the item information functions at that level (Baker and Kim, 2004). Besides, the standard error of measurement (SEM) at a given ability level (θ) is the reciprocal of the square root of the test information (Embretson and Reise, 2000). Thus, it is important to determine at which ability level the TIF peaks for identifying potential gaps in measurement accuracy. To investigate measurement precision, the item information functions (IIFs) for each item in the pilot testing, the TIF for the whole pilot test, and the TIF for each stage in the MST were generated.



4. RESULTS

This chapter presents findings from the development of an assessment framework, as well as the development and administration of a multistage test designed to precisely measure preservice teachers' assessment literacy.

4.1. Results of the Assessment Framework Study

As mentioned in the methodology section, the experts evaluated the draft of the assessment framework developed by the researcher after literature review. The expert meetings resulted in several crucial findings regarding the framework development and significant revisions. To illustrate, in the first draft, the dimensions were named as knowledge-based since the knowledge base of the TALiP (Xu and Brown, 2016) laid the foundation of the draft. However, initial meetings led to turning these knowledge-based dimension such as “The knowledge of grading” to competency-based dimensions such as “Competency in grading”. Both the labels and the scope of the dimensions in the framework were extended to competencies thanks to the discussions with experts. In addition to this, two of the dimension names were changed. “Scoring” was replaced with “Grading” and “Digital assessment” was replaced with “Technology-based assessment”.

Additionally, several objectives were revised based on feedback from subject matter experts. Specifically, multiple experts emphasized the significance of taxonomies and appropriateness to learning objectives within the assessment framework. Consequently, two new objectives were incorporated: “Determines appropriate assessment tools aligned with the intended cognitive level/taxonomy” and “Determines appropriate assessment tools aligned with the intended learning objective” under the first dimension, and “Develops assessment tools aligned with the intended cognitive level/taxonomy” and “Develops assessment tools aligned with the intended learning objective” under the eighteenth dimension. While the initial draft only encompassed self-assessment and peer-assessment as per the TALiP framework (Xu and Brown,

2016), several experts emphasized the significance of incorporating group-based assessment as well. Thus, the review process also led to the inclusion of group assessment as an additional component.

Furthermore, exemplary situations were incorporated into some objectives. For example, several experts expressed uncertainty regarding the content of certain objectives, but they gained a more comprehensive understanding through researchers' illustrative demonstrations. Consequently, it was decided to provide exemplary illustrations for this objective, such as "Knows / Applies the principles of ethical assessment (avoiding sharing exam results with people other than students and their parents, implementing appropriate assessment modifications for students with special needs, not using assessment as a way of punishment, etc.)". Similarly, the seventh dimension elaborated on the concept of technology-based practices by providing several illustrative examples such as e-portfolio.

In addition to providing two illustrative examples, new objectives were incorporated based on experts' recommendations. To illustrate, the framework was revised to include an objective titled "Develops rubrics" under the competency in grading dimension, and an objective titled "Develops appropriate item types according to the assessment purpose" was incorporated under the competency in developing assessment tools dimension. Moreover, the discussions involved some conflicting suggestions from experts. One of these conflicting issues was about the distinction between grading and feedback. One of the experts stated that:

"While grading, our main purpose is actually to give feedback. Thus, these two dimensions are too interrelated. I am not sure how you can separate these. I suggest you consider grading as an objective under the feedback dimension".

While another expert answered by saying:

"Feedback should be assessed separately. Of course, we can give feedback while we also do grading and the purpose of grading is giving feedback. On the other hand, there are also other situations that only one of them occurs. Grading might not lead to feedback when we do not give information about the strengths

and weaknesses resulting in that grade. So, I believe that these two should be considered in two separate dimensions”.

This perspective was further supported by another expert, who argued that “Of course, grading and feedback are two distinct abilities. Feedback can be given without grading just as in formative assessment approaches”. After observing that experts approached this issue from divergent perspectives, the researcher purposefully inquired, “Do you believe feedback and grading competencies should be evaluated independently?” during consecutive meetings. As a result, since the majority of experts concurred that grading may not necessarily yield feedback, and the literature aligns with this perspective, these two dimensions were maintained as distinct elements within the draft version of the framework.

Another dimension that received a lot of discussion was the fifth dimension addressing student involvement in assessment. Some of the experts argued that self-assessment, peer assessment and group assessment methods can be included in the first dimension regarding assessment methods. For instance, one of the experts stated that:

“Of course, these assessment types are significant for both teachers and students. However, I am not sure whether their significance necessitates assessing them in an individual dimension. If I were to develop an assessment framework for assessment literacy, I would likely incorporate these assessment methods as alternative assessment approaches”. The other four specialists also agreed on this approach. However, most experts defend that these assessment methods should be individually assessed. One from this side asserted that “I might prefer to consider these methods under formative assessment. They are mainly used for improving students’ metacognitive skills. Actually, the main aim is not to assess their friends, but rather to assess their own learning process. Thus, I strongly suggest including them in the framework but emphasizing this view and not utilizing the methods for grading purposes. Then, they can be placed as one distinct dimension”..

Another expert also confirmed that by saying “Each of self-assessment, peer assessment and group assessment can be applied in situations that are not performance-based. For these methods, students are the assessors; thus, they have different dynamics”. Besides, another expert emphasized that the MoNE firmly suggests the use of

self-assessment and peer assessment in the new curriculum called the Maarif Model. The expert underlined that although these methods might be considered formative, the curriculum strongly suggests their use regardless of formative or summative approach to assessment.

In addition to these opinions, the newly proposed dimension that was named as “Competency in technology-based assessment” was strengthened with the help of experts’ recommendations. As an example, the phrase of “technology-based assessment methods” was included in the subdimensions. However, one of the experts mentioned that:

“When you say method, then you restrict the scope of the dimension. We might give the example of adaptive tests as such methods, yet there are not many other examples. The use of method word does not involve Web 2.0 tools or other applications that we can consider as technology-based assessment. So, you might revise it as” technology-based assessment practices or applications “to broaden the scope of the dimension”.

This feedback received due consideration and the objectives were adjusted to include “technology-based practices” phrase rather than “methods”. Additionally, the experts cautioned against using specific digital tools when writing items for this dimension, as such tools may become out of use rapidly due to the fast-paced changes in technology.

Consequently, the dimensions in assessment framework for measuring AL was designed as in the following figure in the light of the semi-structured expert meetings and the review of literature. As Figure 4.1 demonstrates, the dimensions were organized non-hierarchically, with each competency dimension maintaining equivalent significance based on its specific requirements.



Figure 4.1. Dimensions of the assessment framework for assessment literacy.

The scope of the dimensions was defined by the objectives. Collectively, the assessment framework was finalized with a total of eight dimensions and fifty-one objectives:

Table 4.1. Teacher assessment literacy framework.

1. Competency in assessment purposes and methods	2. Competency in interpreting and sharing the data obtained from assessment results
--	---

Table 4.1 Teacher assessment literacy framework. (cont.)

· Determines the purpose of assessment · Knows / Applies different types of assessment based on assessment purpose (diagnostic, formative, summative etc.) · Knows / Applies different types of assessment based on the assessment criterion (norm-referenced, criterion-referenced) · Knows / Applies traditional and alternative (performance-based) assessment methods · Determines appropriate assessment tools aligned with the intended cognitive level/taxonomy · Determines appropriate assessment tools aligned with the intended learning objective	· Knows / Applies the fundamental statistical methods required for interpreting assessment results (measures of central tendency, measures of change, standard scores, t score, z score, etc.) · Knows the sources of measurement errors affecting assessment results · Detects measurement error types · Knows the factors affecting the reliability of assessment results · Knows the factors affecting the validity of assessment results · Knows / Applies types of validity evidence · Knows / Applies types of reliability evidence · Shares the data obtained from assessment results effectively to other stakeholders such as students, parents and administrators
3. Competency in grading · Knows the purpose of grading · Knows / Applies the types of criteria used in grading (norm based, criterion based) · Performs effective grading based on the purpose · Weighs the measurement results	4. Competency in feedback · Knows the purpose of feedback · Knows / Applies different types of feedback (formative, evaluative, appreciative etc.) · Knows the advantages and limitations of different feedback methods

Table 4.1 Teacher assessment literacy framework. (cont.)

correctly in the grading process · Knows the types of rubrics (analytical, holistic) · Selects the rubric according to the purpose · Develops rubric	· Gives effective feedback based on the purpose
5. Competency in involving <i>students in assessment</i> · Knows the purpose of self-assessment, peer assessment and group assessment · Knows / Applies the issues to be considered in self-assessment, peer assessment and group assessment · Knows the benefits and limitations of self-assessment, peer assessment and group assessment · Knows / Applies methods for self-assessment, peer assessment and group assessment	6. Competency in assessment ethics · Knows / Applies the principles of ethical assessment (avoiding sharing exam results with people other than students and their parents, implementing appropriate assessment modifications for students with special needs, not using assessment as a way of punishment, etc.) · Knows the types and causes of bias · Knows / Applies the measures to prevent scorer bias · Knows / Applies how to detect and the measures to be taken for biased items that may work in favor of any group · Knows / Applies copy detection methods
7. Competency in technology-based assessment · Selects technology-based assessment applications or tools based on the purpose (web 2.0 tools	8. Competency in developing assessment tools · Knows / Applies the steps or standardized processes for developing assessment tools

Table 4.1 Teacher assessment literacy framework. (cont.)

such as Kahoot, Quizlet, Moodle, VR, e-portfolio etc.) · Uses technology-based assessment applications or tools · Knows various benefits and limitations of technology-based assessment · Knows the properties of adaptive tests · Interprets the results obtained from adaptive tests · Develops questions with technology-supported / artificial intelligence-supported tools · Adapts questions generated by technology-supported / artificial intelligence-supported tools to ensure alignment with learning objectives, validity of assessment outcomes etc.	· Knows/ Applies the principles of developing different types of assessment tools based on the purpose · Adapts existing assessment tools to ensure alignment with learning objectives, assessment standards, validity of assessment outcomes etc.” · Knows the meaning of concepts related to item analysis (item discrimination index, item difficulty index) · Develops assessment tools aligned with the intended cognitive level/taxonomy · Develops assessment tools aligned with the intended learning objective · Knows / Applies types of validity evidence · Knows / Applies types of reliability evidence · Knows / Applies measures for positively affecting validity of assessment results while developing assessment tools · Knows / Applies measures for positively affecting reliability of assessment results while developing assessment tools
---	---

4.2. Results of Pilot Testing

The pilot phase that includes estimating item parameters by 1-PL IRT was crucial for the development of the MST. A total of 105 items were administered across 7 booklets, each containing 20 items, to 1223 participants. The reliability and validity of test results were evaluated.

4.2.1. The Reliability of Pilot Test Results

For reliability analysis, Cronbach Alfa (α), McDonald Omega (ω), and marginal reliability values were calculated. Table 4.2 indicates the calculated Alfa (α) and Omega (ω) values for each booklet.

Table 4.2. Cronbach's alpha and McDonald's omega estimates for each booklet.

Booklet	Sample Size (N)	Cronbach Alfa (α)	McDonald Omega (ω)
1	172	0.36	0.33
2	159	0.46	0.39
3	153	0.43	0.55
4	171	0.35	0.26
5	194	0.43	0.54
6	195	0.59	0.58
7	179	0.30	0.25

In the literature, there are different evaluation criteria for interpreting Cronbach's Alpha and McDonald's Omega estimates. However, a general rule of thumb implies that Alpha values greater than 0.70 are generally considered acceptable (Cicchetti, 1994; George and Mallery, 2003; Nunnally, 1978). Researchers generally interpret McDonald's Omega values using the Alpha approach, as both estimates are conceptually similar and aim to measure internal consistency. An examination of the calculated estimates for the pilot phase indicates that the internal consistency of the test results did not meet the intended level, as the estimates fell below the acceptable threshold (Table 4.1). The multiple matrix booklet design in pilot testing necessitated a heterogeneous structure, thereby reducing the likelihood of observing high correlations

between items measuring distinct constructs. Nonetheless, the overall marginal reliability, which is an important IRT-based measure of test reliability, was evaluated. On the contrary to the CTT-based measures like alpha, marginal reliability coefficient is not directly influenced by the test length (John and Benet-Martínez, 2000). Rather, it provides an average of measurement error estimates across different ability levels, thus it is useful for adaptive test designs. In the current study, the marginal reliability of the pilot test was calculated as 0.95, indicating strong measurement precision across the ability continuum.

4.2.2. The Validity of Pilot Test Results

The current study employed a multifaceted approach to collect content validation evidence, including the use of a table of specifications, expert opinions, and the index of item-objective congruence. The table of specification was created as Table 4.3 illustrates.

Table 4.3. Table of specification.

Booklet	Content Domain (Dimensions)								Cognitive Domain		
	D1	D2	D3	D4	D5	D6	D7	D8	Knowing	Applying	
1	3	3	2	2	2	2	3	3	13 65%	7 35%	20
2	2	2	2	2	3	3	3	3	8 40%	12 60%	20
3	2	3	3	3	2	2	2	3	11 55%	9 45%	20
4	2	3	3	2	2	3	2	3	11 55%	9 45%	20
5	3	3	2	2	2	2	2	4	9 45%	11 55%	20
6	2	3	2	3	2	2	2	4	10 50%	10 50%	20
7	2	3	2	2	3	2	2	4	12 60%	8 40%	20

Additionally, the index of item-objective congruence (IOC), as described by Rovinelli and Hambleton (1977), was evaluated to assess how well an item aligns with the test's objectives, providing aluable information on item validity (Crocker and Algina, 2006). Ratings from seven experts were used to calculate the IOC index for each item (Appendix A). The calculations revealed that the IOC indices ($N = 105$) ranged from -0.29 to 1.00. It is recommended that items with IOC index less than 0.75 should be revised or removed through careful inspection (Tuner and Carlson, 2003). In the current study, most indices (64.8%) were found to be equal to or greater than 0.75. Thirty-nine out of the 105 items were found to have an IOC index value of less than 0.75. These items were evaluated and modified if necessary.

Construct validity was examined using the pilot test results. Initially, an Exploratory Factor Analysis (EFA) with Principal Axis Factoring extraction and oblimin rotation was conducted to investigate the underlying factor structure. The Bartlett's Test of Sphericity was significant ($p < 0.05$) for five out of the seven booklets, suggesting adequate correlations for factor analysis (Bartlett, 1950). However, the test failed to reach significance for Booklet 1 and Booklet 7. The Kaiser-Meyer-Olkin measure displayed marginal-to-adequate values ranging from 0.49 to 0.64 across the booklets (Kaiser, 1960). Notably, the KMO values fell below the recommended threshold of 0.60 for six of the booklets, indicating potential issues with sampling adequacy or item covariance.

Moreover, The EFA results revealed suboptimal factor structures across all booklets, with explained variance ratios ranging from 29.95% to 37.38%. This limited variance explanation, coupled with several items exhibiting weak factor loadings and high uniqueness. The summary of EFA results for each booklet was shown in Table 4.4. (see Appendix B for full results).

Table 4.4. Summary of EFA results across seven booklets.

Booklet	N	KMO	Bartlett's Test (p)	Factors Retained (Eigen- values >1)	Total Variance Expla- ined (%)	Item Com- Muna- lities Range	Factor Loadings Range
1	172	0.50	0.30	10	31.96	0.13- 0.64	NA
2	159	0.51	0.00	9	34.75	0.10- 0.76	-0.69- 0.87
3	153	0.55	0.00	8	31.52	0.11- 0.70	-0.67- 0.85
4	171	0.55	0.05	9	30.62	0.03- 0.68	-0.61- 0.80
5	194	0.51	0.00	10	37.38	0.11- 0.70	-0.80- 0.83
6	195	0.64	0.00	8	29.95	0.04- 0.55	NA
7	179	0.49	0.13	10	32.64	0.07- 0.68	NA

Following EFA, a confirmatory factor analysis (CFA) was performed for each booklet. However, the CFA conducted in JASP (JASP Team, 2023) was unable to generate a viable model for the booklet data, resulting in an error due to negative variance estimates for some variables as shown in Appendix C.

4.2.3. Results of Item Response Theory Analysis

Before proceeding with further item analysis, the assumptions underlying the IRT were rigorously evaluated. In the current study, the monotonicity assumption was assessed through Item Characteristic Curves (ICCs), which are expected to accurately represent the relationship between the latent trait and item responses (Baker, 2001). Figure 4.2. presents the ICCs for each of the 105 items. As the ability level increases, the probability of answering each item correctly also increases monotonically. Consequently, smooth S-shaped curves were generated as anticipated, thereby confirming that the monotonicity assumption is upheld.

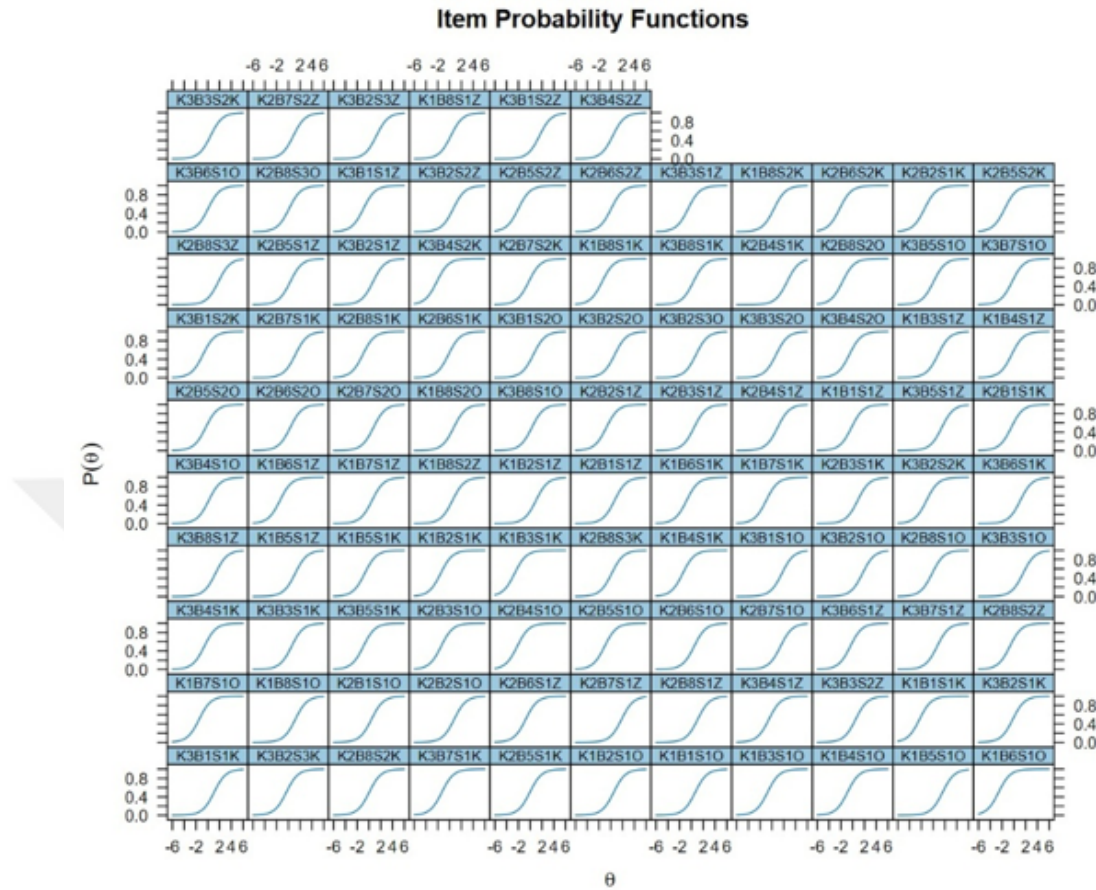


Figure 4.2. Item characteristic curves for pilot test items.

The other assumption is local independence which implies that an examinee's probability of answering an item correctly is independent of their response behavior to other items on the test (Baker, 2001). To test local independence, Yen's Q3 Statistic test was utilized. The Yen's Q3 test results indicate that the assumption of local independence is not violated, as all Q3 values are greater than 0.37, suggesting a high level of independence among the items (Smits *et al.*, 2012). However, unidimensionality assumption was violated according to the Kaiser rule, parallel analysis (Horn, 1965; Green *et al.*, 2012), and the first-to-second eigen value ratio approach (Lord, 1980). The first-to-second eigen value ratios for each booklet were found to be 1.14, 1.22, 1.13, 1.56, and 1.06 respectively.

To evaluate the fit of the items to the model, Orlando and Thissen's Chi-Square Item Fit values, also known as the S-X2 statistic, were computed (Orlando and Thissen, 2000). The results are indicated in Table 4.5. S-X2 values ranged from 2.83 to 38.34. Based on the calculated item fit values, most items showed a good fit to the 1-PL model ($p > 0.05$, RMSEA < 0.06). In line with Hu and Bentler's (1999) suggestion, items demonstrating RMSEA values below 0.06 were considered to have acceptable model-data fit, and all RMSEA values are found to be as less than 0.06. Nevertheless, despite achieving these RMSEA criteria, certain items, such as K3B4S2Z and K2B1S1O, exhibited indications of misfit, as their significance levels did not meet the predetermined thresholds ($p = 0.01$, ≤ 0.05). Consequently, these items were identified as potentially misfitting and warranting further review for possible revision or consideration for inclusion in the MST item bank.

Table 4.5. Chi-square item fit values for pilot test.

Item	S-X2	df	RMSEA	p
K3B1S1K	14.30	9.00	0.02	0.11
K3B2S3K	14.59	9.00	0.02	0.10
K2B8S2K	10.32	9.00	0.01	0.33
K3B7S1K	16.55	9.00	0.03	0.06
K2B5S1K	14.84	9.00	0.02	0.10
K1B2S1O	19.47	9.00	0.03	0.02
K1B1S1O	20.96	9.00	0.03	0.01
K1B3S1O	23.51	9.00	0.04	0.01
K1B4S1O	32.94	9.00	0.05	0.00
K1B5S1O	6.12	9.00	0.00	0.73
K1B6S1O	10.77	9.00	0.01	0.29
K1B7S1O	27.16	9.00	0.04	0.00
K1B8S1O	13.64	9.00	0.02	0.14

Table 4.5 Chi-square item fit values for pilot test. (cont.)

Item	S-X2	df	RMSEA	p
K2B1S1O	23.42	9.00	0.04	0.01
K2B2S1O	14.82	9.00	0.02	0.10
K2B6S1Z	14.87	9.00	0.02	0.09
K2B7S1Z	5.96	9.00	0.00	0.74
K2B8S1Z	2.83	9.00	0.00	0.97
K3B4S1Z	8.77	9.00	0.00	0.46
K3B3S2Z	12.55	9.00	0.02	0.18
K1B1S1K	17.36	9.00	0.03	0.04
K3B2S1K	20.75	9.00	0.03	0.01
K3B4S1K	5.78	9.00	0.00	0.76
K3B3S1K	14.25	9.00	0.02	0.11
K3B5S1K	8.33	9.00	0.00	0.50
K2B3S1O	35.83	9.00	0.05	0.00
K2B4S1O	21.18	9.00	0.03	0.01
K2B5S1O	13.18	9.00	0.02	0.15
K2B6S1O	9.04	9.00	0.00	0.43
K2B7S1O	26.98	9.00	0.04	0.00
K3B6S1Z	18.87	9.00	0.03	0.03
K3B7S1Z	13.42	9.00	0.02	0.14
K2B8S2Z	23.70	9.00	0.04	0.00
K3B8S1Z	8.47	9.00	0.00	0.49
K1B5S1Z	9.24	9.00	0.00	0.42
K1B5S1K	18.64	9.00	0.03	0.03
K1B2S1K	17.56	9.00	0.03	0.04
K1B3S1K	12.14	9.00	0.02	0.21
K2B8S3K	13.51	9.00	0.02	0.14
K1B4S1K	9.95	9.00	0.01	0.35
K3B1S1O	8.41	9.00	0.00	0.49
K3B2S1O	10.18	9.00	0.01	0.34

Table 4.5 Chi-square item fit values for pilot test. (cont.)

Item	S-X2	df	RMSEA	p
K2B8S1O	33.92	9.00	0.05	0.00
K3B3S1O	9.07	9.00	0.00	0.43
K3B4S1O	16.31	9.00	0.03	0.06
K1B6S1Z	20.68	9.00	0.03	0.01
K1B7S1Z	9.44	9.00	0.01	0.40
K1B8S2Z	20.81	9.00	0.03	0.01
K1B2S1Z	21.23	9.00	0.03	0.01
K2B1S1Z	7.80	9.00	0.00	0.55
K1B6S1K	9.14	9.00	0.00	0.42
K1B7S1K	14.48	9.00	0.02	0.11
K2B3S1K	33.46	9.00	0.05	0.00
K3B2S2K	19.78	9.00	0.03	0.02
K3B6S1K	15.00	9.00	0.02	0.09
K2B5S2O	38.34	9.00	0.05	0.00
K2B6S2O	14.41	9.00	0.02	0.11
K2B7S2O	28.29	9.00	0.04	0.00
K1B8S2O	26.28	9.00	0.04	0.00
K3B8S1O	18.88	9.00	0.03	0.03
K2B2S1Z	9.05	9.00	0.00	0.43
K2B3S1Z	8.33	9.00	0.00	0.50
K2B4S1Z	14.24	9.00	0.02	0.11
K1B1S1Z	16.52	9.00	0.03	0.06
K3B5S1Z	8.51	9.00	0.00	0.48
K2B1S1K	17.63	9.00	0.03	0.04
K3B1S2K	15.59	9.00	0.02	0.08
K2B7S1K	15.88	9.00	0.03	0.07
K2B8S1K	12.62	9.00	0.02	0.18
K2B6S1K	4.57	9.00	0.00	0.87

Table 4.5 Chi-square item fit values for pilot test. (cont.)

Item	S-X2	df	RMSEA	p
K3B1S2O	32.42	9.00	0.05	0.00
K3B2S2O	28.50	9.00	0.04	0.00
K3B2S3O	21.22	9.00	0.03	0.01
K3B3S2O	15.49	9.00	0.02	0.08
K3B4S2O	11.86	9.00	0.02	0.22
K1B3S1Z	17.09	9.00	0.03	0.05
K1B4S1Z	16.81	9.00	0.03	0.05
K2B8S3Z	14.04	9.00	0.02	0.12
K2B5S1Z	17.42	9.00	0.03	0.04
K3B2S1Z	10.93	9.00	0.01	0.28
K3B4S2K	18.71	9.00	0.03	0.03
K2B7S2K	14.28	9.00	0.02	0.11
K1B8S1K	7.04	9.00	0.00	0.63
K3B8S1K	19.85	9.00	0.03	0.02
K2B4S1K	9.40	9.00	0.01	0.40
K2B8S2O	19.72	9.00	0.03	0.02
K3B5S1O	17.05	9.00	0.03	0.05
K3B7S1O	13.72	9.00	0.02	0.13
K3B6S1O	23.24	9.00	0.04	0.01
K2B8S3O	37.91	9.00	0.05	0.00
K3B1S1Z	24.06	9.00	0.04	0.00
K3B2S2Z	32.93	9.00	0.05	0.00
K2B5S2Z	10.82	9.00	0.01	0.29
K2B6S2Z	12.42	9.00	0.02	0.19
K3B3S1Z	10.34	9.00	0.01	0.32
K1B8S2K	6.01	9.00	0.00	0.74
K2B6S2K	15.66	9.00	0.02	0.07
K2B2S1K	18.23	9.00	0.03	0.03

Table 4.5 Chi-square item fit values for pilot test. (cont.)

Item	S-X2	df	RMSEA	p
K2B5S2K	12.12	9.00	0.02	0.21
K3B3S2K	14.83	9.00	0.02	0.10
K2B7S2Z	8.42	9.00	0.00	0.49
K3B2S3Z	12.91	9.00	0.02	0.17
K1B8S1Z	12.62	9.00	0.02	0.18
K3B1S2Z	12.32	9.00	0.02	0.20
K3B4S2Z	22.31	9.00	0.03	0.01

According to the generated 1-PL model, b parameters range from -2.64 to 2.57, with a mean of 0.02 (SD = 1.10). Table 4.5. demonstrates descriptive information for b values. The wide range of b-parameter values suggests that the test has potential to effectively differentiate between individuals at both low and high ability levels, indicating a broad scope of measurement (Figure 4.3).

Table 4.6. Descriptive statistics of the item difficulty parameters.

Statistic	N	Min	Max	Mean	Median	Standard Deviation	Skewness	Kurtosis
b-parameter	105	-2.64	2.57	-0.02	0.16	1.10	-0.33	-0.39

Consequently, the IRT analysis of pilot test items under the one-parameter logistic model revealed two key patterns that informed the development of the MST: the item difficulty parameters spanned a broad continuum, and the item fit statistics identified potential misfitting items.

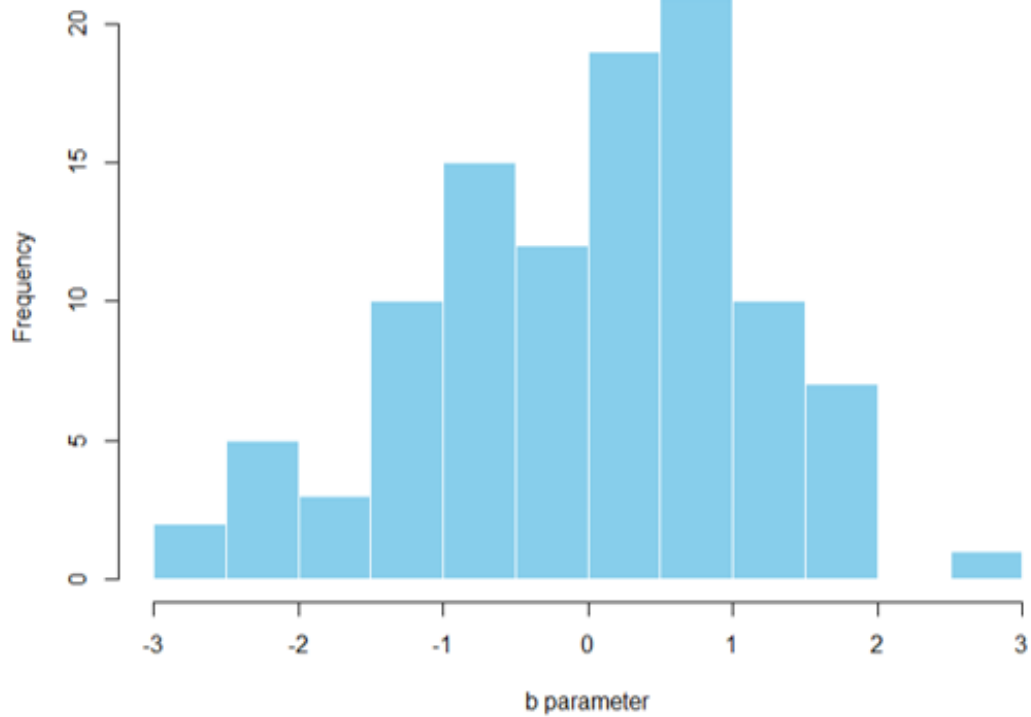


Figure 4.3. Histogram of item difficulty parameters.

Moreover, the test information function (TIF) and standard error of measurement (SEM) were analyzed and graphically represented. As illustrated in Figure 4.4, the TIF peaks at approximately $I(\theta) \approx 20$ near $\theta = 0$, indicating that the test provides the highest precision for individuals with average ability levels. The standard error exhibits an inverse relationship with the TIF consistent with psychometric theory. The SEM reaches its minimum value ($SEM(\theta) \approx 0.2$) around $\theta = 0$, reflecting maximal measurement accuracy within this range. Furthermore, the test items demonstrate substantial informational coverage across a broad spectrum of abilities, yielding meaningful data for participants with ability levels ranging from $\theta = -3$ to $\theta = +3$.

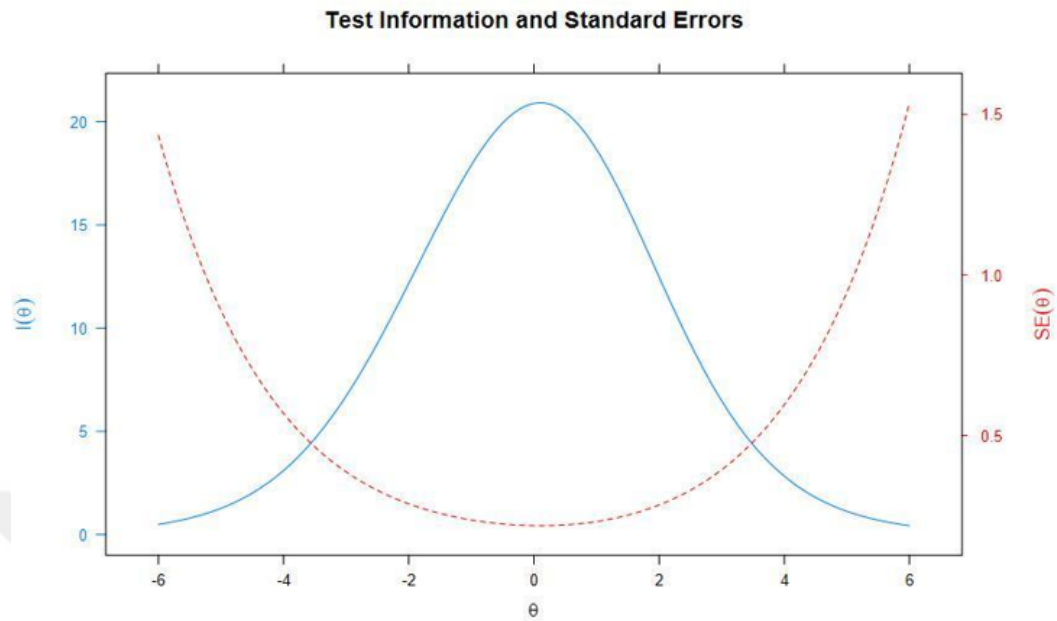


Figure 4.4. Test information function and SEM across ability levels.

Additionally, the item information functions were examined for each item. The item information function plays a crucial role in test development, as it reveals how effectively each test item contributes to the estimation of examinees' ability levels. (Hambleton *et al.*, 1985). The generated function suggests varying sensitivity across different ability levels, reflecting differential measurement precision for examinees with diverse proficiency ranges (Figure 4.5).

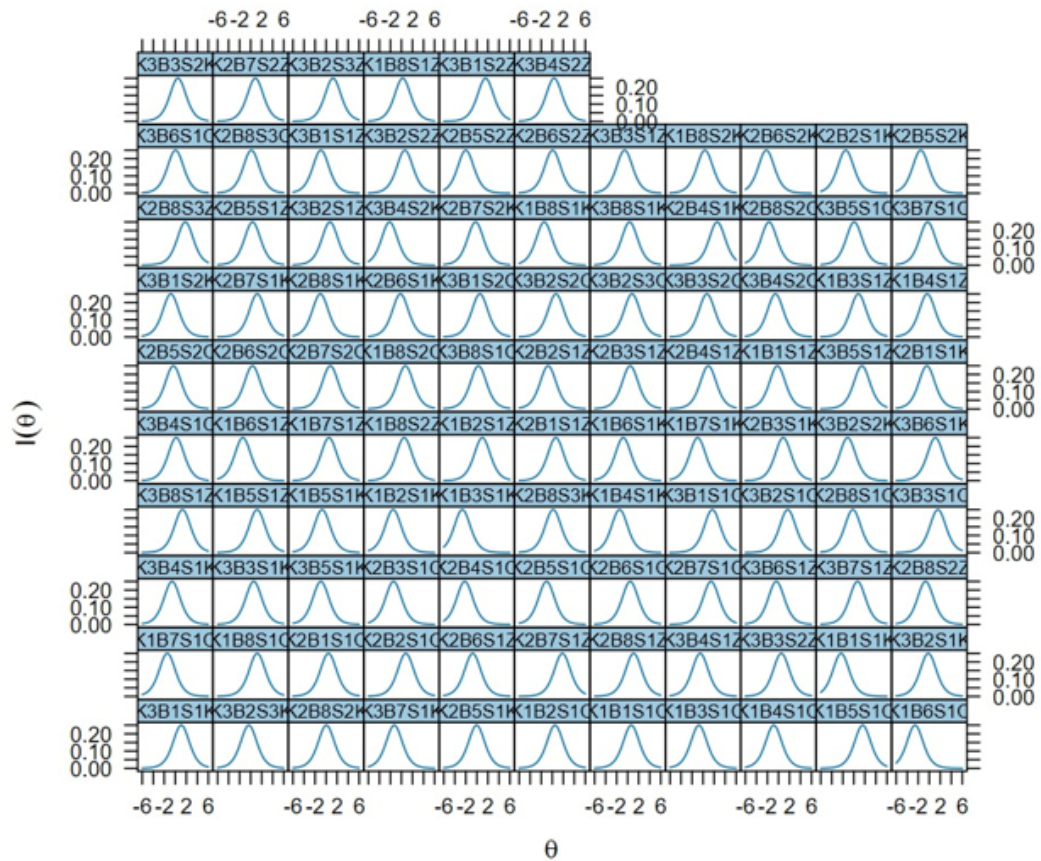


Figure 4.5. Item information functions for pilot test items.

4.3. Results of the MST

Drawing insights from the pilot testing findings, the MST consists of 3 stages and 7 modules was created. When selecting the specific items from the pool of 105 items to include in the MST, the item difficulty (b) parameters and item-fit statistics were carefully considered. The routing module was constructed using medium-difficulty items. For subsequent modules, careful attention was paid to ensure that items belonging to the same subdimension represented varying difficulty levels (b-parameters) that could be classified as easy, medium, and hard. The b-parameters for items included in the test modules are presented in Table 4.7.

Table 4.7. Distribution of item difficulty parameters for MST modules.

Routing Module (Medium)					
Item	b				
K1B1S1O	0.66				
K2B2S1Z	- 0.75				
K3B3S2Z	0.52				
K3B4S2Z	0.34				
K2B5S1Z	0.34				
K1B6S1K	0.33				
<i>Mean</i>	0.15				
Second Stage (Easy)		Second Stage (Medium)		Second Stage (Hard)	
Item	b	Item	b	Item	b
K2B1S1K	-0.43	K2B1S1O	0.43	K2B1S1Z	0.64
K3B2S1K	-0.17	K3B2S1Z	0.7	K3B2S1O	1.68
K1B3S1K	-2.64	K1B3S1O	-0.75	K1B3S1Z	0.82
K1B4S1K	-1.46	K1B4S1O	-0.86	K1B4S1Z	-0.36
K1B5S1K	-0.72	K1B5S1Z	1.17	K1B5S1O	1.6
K2B6S1Z	-0.76	K2B6S1K	-0.19	K2B6S1O	0.47
<i>Mean</i>	-1.12	<i>Mean</i>	0.05	<i>Mean</i>	0.84
Third Stage (Easy)		Third Stage (Medium)		Third Stage (Hard)	
Item	b	Item	b	Item	b
K3B1S2K	-0.81	K3B1S2O	0.54	K3B1S2Z	1.58
K3B2S3K	-0.34	K3B2S3O	0.79	K3B2S3Z	1.26
K3B3S1Z	-0.55	K3B3S1K	0.55	K3B3S1O	1.57
K2B4S1O	-2.2	K2B4S1Z	0.95	K2B4S1K	2.57

Table 4.7 Distribution of item difficulty parameters for MST modules. (cont.)

Third Stage (Easy)		Third Stage (Medium)		Third Stage (Hard)	
K3B5S1K	-0.95	K3B5S1O	0.06	K3B5S1Z	1.46
K3B6S1Z	-0.39	K3B6S1O	0.05	K3B6S1K	1.11
K2B7S2K	-0.24	K2B7S1O	1.45	K2B7S2Z	0.85
<i>Mean</i>	-0.81	<i>Mean</i>	0.53	<i>Mean</i>	1.44

For instance, the initial items in Stage 2 (coded as K2B1S1K, K2B1S1O, and K2B1S1Z) share identical content scope and dimensionality but are differentiated by their difficulty parameters. This strategic variation in item difficulty allows for more precise measurement across the ability spectrum while maintaining content consistency.

The various findings were obtained through the administration of the MST to 106 participants. To illustrate, the paths that were followed by the participants were recorded (Figure 4.6). After completing the routing module in the first stage, the majority of participants were most frequently routed to the difficult module in the second stage (53), followed by the medium module (48), with only a small number directed to the easy module (5). The majority of participants who started with the medium module either (a) continued through medium-difficulty modules, finishing the test at this level ($n = 39$), or (b) advanced to difficult modules for test completion ($n = 39$), according to an analysis of routing decisions across MST stages. In contrast, the least frequent trajectory was used in both succeeding stages, with only two individuals being directed solely to easy modules.

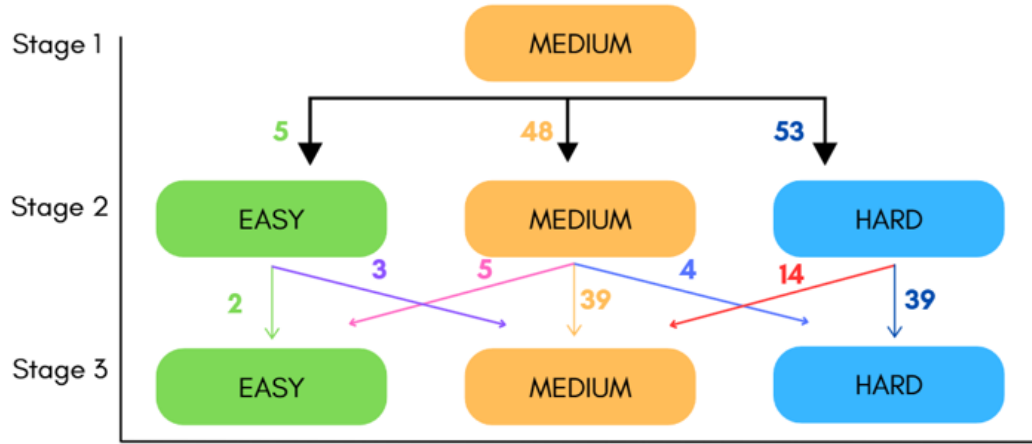


Figure 4.6. Distribution of routing module selections by participants.

Moreover, Table 4.8 demonstrates the ability level (θ) and SEM values for each participant after each test stage. As illustrated in Table 1.1, the column θ_1 represents participants' estimated ability levels after completing the first test stage (medium-difficulty module), while SEM_1 denotes the corresponding standard error of measurement for this stage. Similarly, θ_2 reflects ability estimates following the second stage, with SEM_3 indicating its associated standard error. As expected, participants' θ values should theoretically increase with each successive test stage. Thus, improved measurement precision was empirically demonstrated by the declining standard error values (Mean $SEM_1 = 0.43$, Mean $SEM_2 = 0.32$, Mean $SEM_3 = 0.27$). This gradual decrease in measurement error shows increased accuracy in determining individual skill levels as well as enhanced the reliability of test results. Besides, the MST system successfully produced accurate ability estimates for each participant while preserving low measurement error throughout the adaptive stages, as evidenced by the fact that most participants finished the test with standard errors less than 0.30.

Table 4.8. Ability estimates and standard errors in MST.

θ_1	SEM_1	θ_2	SEM_2	θ_3	SEM_3
1.52	0.56	1.35	0.35	1.51	0.29
0.15	0.41	0.79	0.33	0.89	0.28

Table 4.8 Ability estimates and standard errors in MST. (cont.)

θ_1	SEM_1	θ_2	SEM_2	θ_3	SEM_3
0.74	0.43	0.57	0.34	0.36	0.30
0.44	0.41	0.68	0.33	0.92	0.29
0.15	0.41	0.27	0.31	0.70	0.27
0.15	0.41	0.11	0.31	0.27	0.26
0.15	0.41	-0.23	0.32	-0.09	0.27
0.15	0.41	0.04	0.32	0.20	0.26
0.94	0.49	0.71	0.33	0.75	0.28
0.44	0.41	0.46	0.31	0.52	0.27
0.74	0.43	0.29	0.32	0.40	0.26
0.44	0.41	0.48	0.32	0.33	0.29
0.15	0.41	-0.02	0.32	-0.05	0.27
0.74	0.43	0.48	0.32	0.61	0.28
-0.15	0.43	-0.23	0.32	-0.11	0.27
0.15	0.41	0.27	0.31	0.35	0.26
0.44	0.41	0.11	0.32	0.36	0.27
-0.15	0.43	0.27	0.31	0.27	0.26
0.44	0.41	0.29	0.32	0.40	0.26
-0.15	0.43	0.11	0.31	0.22	0.27
-0.15	0.43	0.27	0.31	0.35	0.26
1.52	0.56	0.76	0.33	0.50	0.29
-0.15	0.43	0.12	0.34	0.30	0.28
-0.15	0.43	-0.61	0.34	-0.54	0.27
-0.15	0.43	-0.02	0.32	0.19	0.27
0.74	0.43	1.01	0.34	1.03	0.27
1.08	0.47	0.62	0.31	0.62	0.28
0.44	0.41	0.62	0.31	0.81	0.27
0.74	0.43	0.79	0.32	0.58	0.28
-0.15	0.43	-0.06	0.31	0.01	0.26
0.15	0.41	0.22	0.32	0.60	0.27
0.74	0.43	0.62	0.31	0.64	0.27
-0.42	0.47	-0.38	0.35	-0.09	0.30

Table 4.8 Ability estimates and standard errors in MST. (cont.)

θ_1	SEM_1	θ_2	SEM_2	θ_3	SEM_3
0.26	0.44	-0.38	0.34	-0.20	0.27
-0.15	0.43	-0.23	0.32	0.01	0.26
-0.15	0.43	0.15	0.32	0.19	0.27
0.74	0.43	0.79	0.32	0.76	0.26
0.74	0.43	0.65	0.32	0.57	0.28
-0.15	0.43	0.11	0.31	0.46	0.26
0.44	0.41	0.62	0.31	0.43	0.28
-0.42	0.47	-0.76	0.35	-0.38	0.27
-0.15	0.43	0.11	0.31	-0.03	0.27
0.74	0.43	1.58	0.38	1.37	0.28
-0.15	0.43	-0.61	0.34	-0.54	0.27
0.15	0.41	0.27	0.31	0.35	0.26
-0.07	0.45	-0.01	0.32	0.15	0.28
0.15	0.41	-0.14	0.33	-0.38	0.30
0.44	0.41	0.44	0.33	0.45	0.29
0.74	0.43	0.29	0.32	0.25	0.26
0.44	0.41	0.62	0.31	0.60	0.28
-0.15	0.43	-0.23	0.32	-0.31	0.28
0.74	0.43	0.29	0.32	0.28	0.27
0.44	0.41	0.29	0.32	0.25	0.26
0.60	0.45	0.22	0.34	0.05	0.29
0.40	0.42	-0.11	0.35	0.15	0.28
0.57	0.46	0.60	0.35	0.56	0.30
-0.07	0.45	-0.19	0.33	0.02	0.28
0.44	0.41	0.11	0.32	0.25	0.26
0.15	0.41	0.27	0.31	0.35	0.26
0.15	0.41	0.11	0.31	0.15	0.26
0.44	0.41	0.54	0.33	0.42	0.31

Table 4.8 Ability estimates and standard errors in MST. (cont.)

θ_1	SEM_1	θ_2	SEM_2	θ_3	SEM_3
-0.15	0.43	-0.41	0.33	-0.54	0.27
-0.15	0.43	-0.06	0.31	0.13	0.26
0.15	0.41	0.11	0.31	-0.09	0.27
0.60	0.45	0.46	0.34	0.76	0.30
0.15	0.41	0.11	0.31	0.15	0.26
-0.15	0.43	-0.82	0.36	-0.80	0.28
0.56	0.46	0.71	0.33	0.61	0.28
-0.15	0.43	-0.06	0.31	-0.09	0.28
0.44	0.41	0.46	0.31	0.76	0.26
-0.15	0.43	-0.23	0.32	-0.23	0.27
0.15	0.41	0.11	0.31	0.13	0.26
0.44	0.41	0.46	0.31	0.39	0.27
0.15	0.41	0.44	0.31	0.51	0.27
0.74	0.43	0.46	0.31	0.43	0.28
0.44	0.41	0.30	0.32	0.24	0.28
0.15	0.41	0.11	0.31	0.23	0.27
-0.15	0.43	0.07	0.34	0.07	0.29
0.44	0.41	0.29	0.32	0.01	0.27
0.15	0.41	0.44	0.31	0.51	0.27
0.44	0.41	0.29	0.32	0.48	0.26
0.44	0.41	0.96	0.32	0.64	0.27
0.44	0.41	0.29	0.32	0.40	0.26
0.15	0.41	-0.23	0.32	0.01	0.26
0.44	0.41	0.62	0.31	0.64	0.27
1.08	0.47	1.35	0.35	1.00	0.27
0.15	0.41	0.11	0.31	0.13	0.26
1.08	0.47	0.96	0.32	1.25	0.27

Table 4.8 Ability estimates and standard errors in MST. (cont.)

θ_1	SEM_1	θ_2	SEM_2	θ_3	SEM_3
0.74	0.43	0.62	0.31	0.43	0.28
0.44	0.41	0.30	0.32	0.42	0.27
-0.15	0.43	-0.41	0.33	-0.17	0.27
-0.48	0.46	-0.20	0.34	-0.05	0.28
0.44	0.41	0.79	0.32	0.88	0.26
0.44	0.41	0.62	0.31	0.64	0.27
-0.40	0.47	-0.96	0.35	-0.76	0.28
0.26	0.44	0.06	0.35	0.00	0.31
0.15	0.41	0.44	0.31	0.38	0.27
1.08	0.47	0.89	0.34	1.15	0.29
0.26	0.44	-0.01	0.32	-0.08	0.27
0.62	0.45	0.77	0.33	0.65	0.28
-0.15	0.43	0.11	0.31	0.13	0.26
1.52	0.56	0.96	0.32	0.64	0.27
0.74	0.43	0.65	0.32	0.83	0.28
1.52	0.56	0.79	0.32	0.52	0.27
1.52	0.56	0.96	0.32	0.76	0.26
-0.48	0.46	-0.39	0.33	-0.32	0.29

Furthermore, the TIF for each stage in MST was created. As shown in Figure 4.7, the TIF for stage 1 peaks at near $\theta = 0$, typically represents the mean ability level of the population. Thus, the TIF created for the first stage represents that the routing module provides maximum measurement precision for examinees of average ability, ensuring reliable initial classification into subsequent adaptive stages.

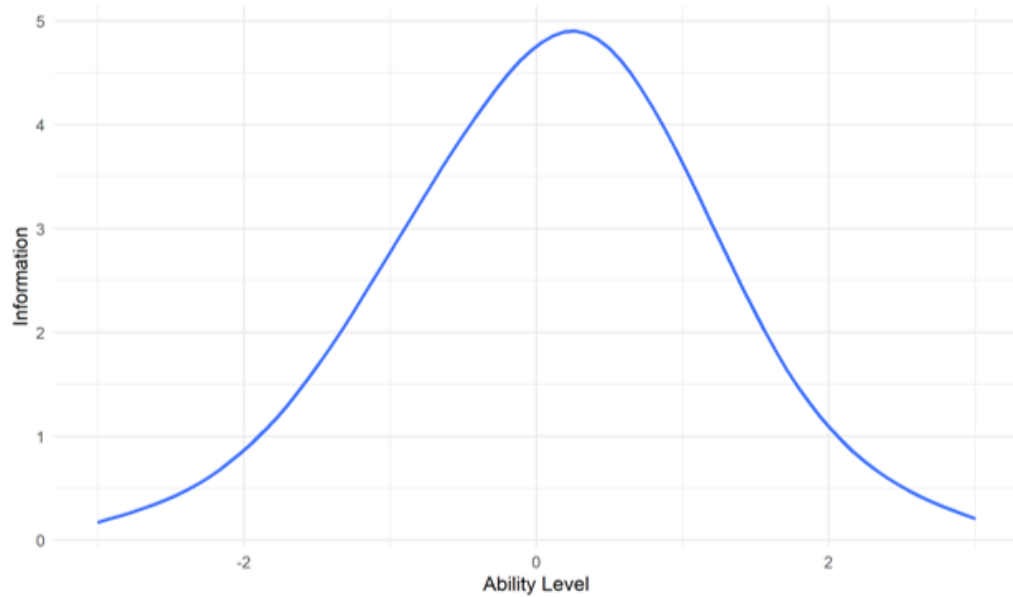


Figure 4.7. The TIF for stage 1.

The TIFs generated for each module in Stage 2 are illustrated in Figure 4.8. According to these TIFs, the easy module peaks around $\theta \approx -1.5$, suggesting it provides the most information for low-ability examinees. The medium difficulty module peaks around $\theta \approx 0.3$ – 0.5 , implying it is optimal for average-ability participants. Lastly, the hard level module peaks around $\theta \approx 1.5$, making it well-suited for high-ability examinees. Similarly, the TIFs generated for Stage 3 (Figure 4.9) demonstrate that each module covers a distinct theta region with good separation and minimal overlap at extremes. At this stage, the easy module peaks around $\theta \approx -1.5$, the medium module peaks near $\theta \approx 0.5$, and the hard module peaks at $\theta \approx 1.7$ – 2.0 . Consequently, the TIFs in Stage 3 peaked higher than those in Stage 2, indicating higher information and greater precision in theta estimation at this stage.

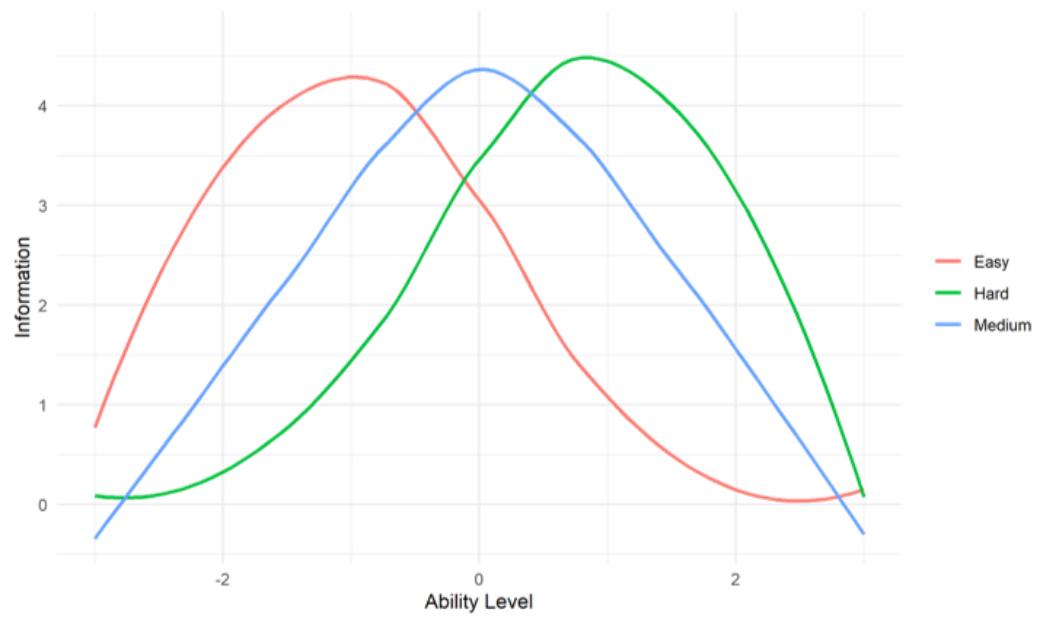


Figure 4.8. The TIF for stage 2.

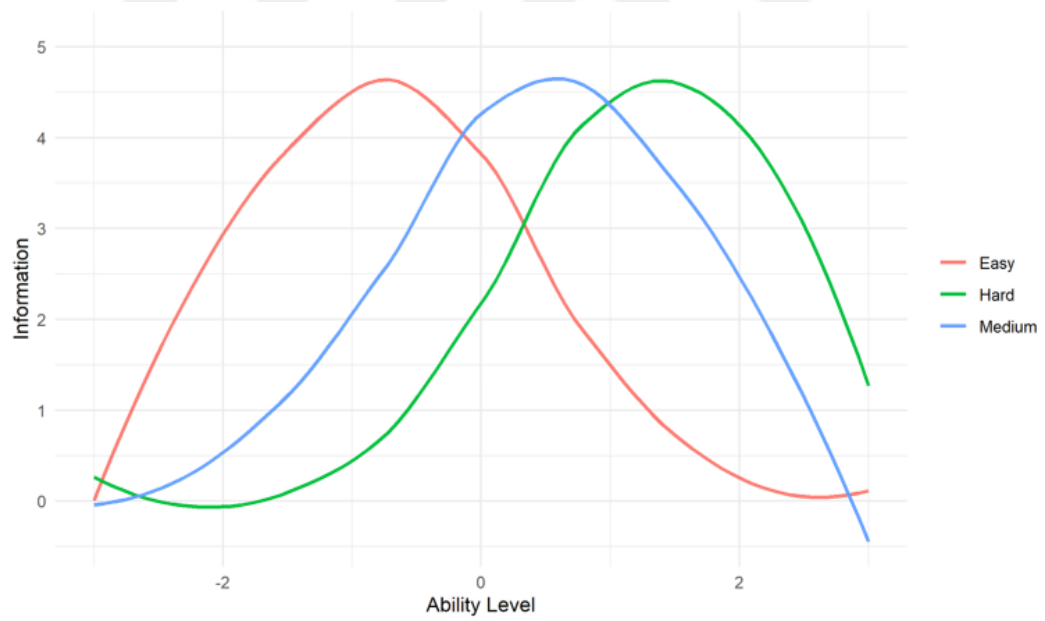


Figure 4.9. The TIF for stage 3.

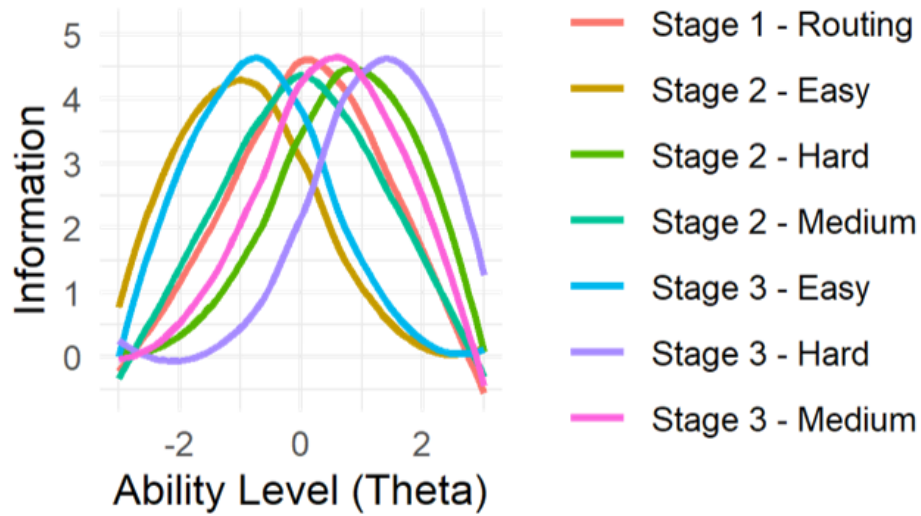


Figure 4.10. The TIFs for all modules.

4.3.1. The Reliability of MST Results

The reliability of the MST results was examined using two different methods. According to the expected a posteriori reliability, which was calculated based on pre-calibrated item parameters, the reliability coefficient of the MST was found to be 0.95, indicating excellent reliability. Additionally, the marginal reliability for each module was computed. The following table shows the marginal reliability for each module, and the results suggested that all calculated marginal reliability coefficients exhibited acceptable levels of reliability (>0.70).

Table 4.9. Marginal reliability for each module in the MST.

Module	Marginal Reliability
Stage 1- Routing	0.78
Stage 2- Easy	0.73
Stage 2- Medium	0.78
Stage 2- Hard	0.75
Stage 3- Easy	0.76
Stage 3- Medium	0.78
Stage 3- Hard	0.70

Additionally, the standard error of ability estimation for each participant was examined to provide further evidence the test's reliability. As the standard error of ability estimation decreases, the reliability of the test increases, indicating that the MST becomes more consistent in estimating the participant's true ability across its stages (Figure 4.11).

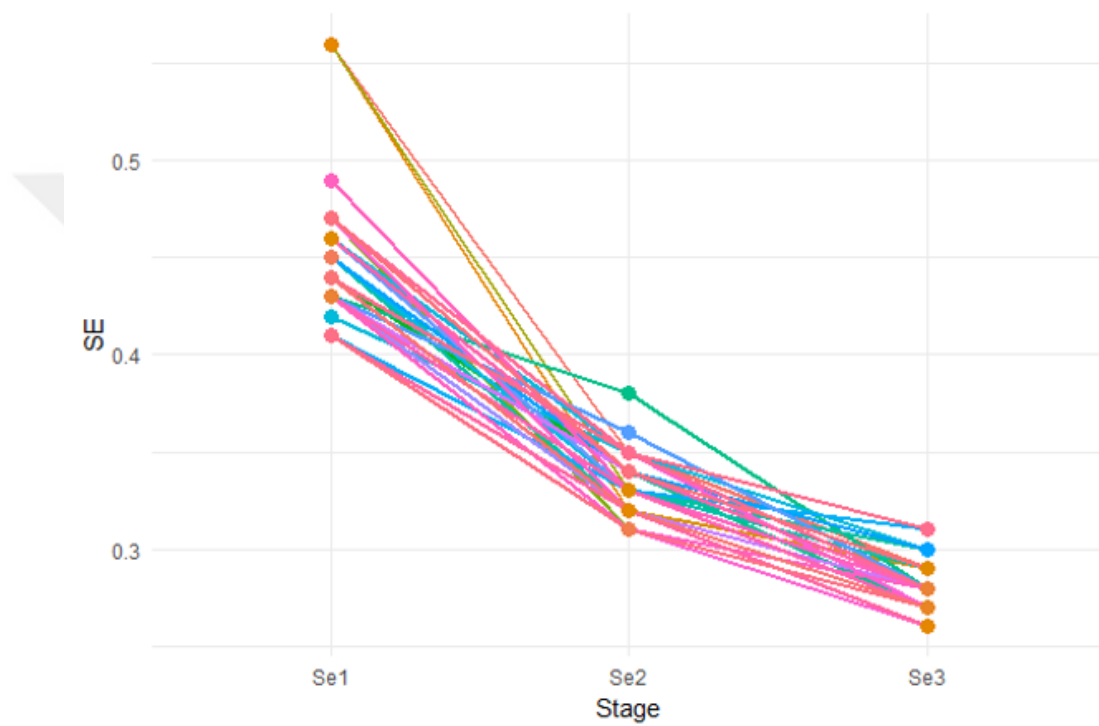


Figure 4.11. Standard errors of ability estimation across stages.

5. DISCUSSION

This chapter includes the key findings and implications from the development of a comprehensive assessment framework and a multistage test for measuring preservice teachers' assessment literacy. The discussion explores each research question, highlighting the progress made in the current study, and outlines the promotion of assessment literacy.

5.1. Discussion of RQ1: Core Dimensions of Assessment Literacy

The current research aimed to extend an existing conceptual framework for assessment literacy to investigate the first research question: "What are the core dimensions of assessment literacy that align with current demands and practices of educational contexts?". As discussed in the literature review, existing measurement instruments are commonly based on the outdated assessment standards. Additionally, although there are several existing frameworks that have potential to inform teacher education or policies for promoting AL, none of the frameworks were specifically designed to facilitate the development of new measurement tools. Thus, extending an existing framework to an assessment framework for AL was a novel approach in the field by this study. On the contrary to Starck *et al.* (2018)'s framework, the current assessment framework was not limited to a specific subject area. Besides, although the teaching competencies developed by the MoNE were considered during the process, the scope was not restricted to these competencies as Alzhrany (2022)'s study in which the alignment of national standards was a crucial concern.

Another significant difference of the extended assessment framework is that it does not include abilities or constructs that are complex, multidimensional, and difficult to assess using a multiple-choice test. For instance, Chan and Luk (2022)'s holistic framework involves the dimensions of attitude and socio-emotional management. It is well-established that the concept of assessment literacy is not only compromised

of a static body of knowledge but rather should be viewed as an evolving construct that develops through experience, professional development, and reflective practice as Xu and Brown (2016) suggests. However, such constructs that might not be directly observable, and can be inferred from behavior were not taken into account by purpose, as the current study aimed to create an assessment framework which might serve as guidance for test developers. Nevertheless, the extended framework did not merely include a set of knowledge related AL, rather create competency-based dimensions to highlight the fact that having required knowledge might not be sufficient for the practical reflection of AL.

Furthermore, the majority of subdimensions were established to consider the contextual factors relevant to assessment literacy. The assessment framework was designed to be flexible, incorporating a range of current practices for assessment literacy to allow for a more sustainable approach. For instance, one of the subdimensions under the “Competency in Assessment Ethics” dimension did not define the principles of ethical assessment with a strict or inflexible approach. Instead, the objective provided several scenarios for such principles, such as not using assessment as a form of punishment. Another example might be preferring the phrases “different assessment methods” or “technology-based assessment applications or tools” rather than limiting the scope of the dimension to methods that today’s teachers generally use, as new assessment methods or tools may be developed to address evolving needs and opportunities in educational assessment. Thus, it might be asserted that the evolving context of educational assessment was considered during the development process.

Notably, while an existing framework, named the TALiP (Xu and Brown, 2016), informed the development process and articulated the conceptual framework for the assessment framework, the present study’s findings contributed to the literature by proposing two novel dimensions. First, previous frameworks do not include a separate dimension that highlights the role of developing different assessment tools. However, the dimension named as “Competency in developing assessment tools” involves several crucial knowledge and abilities that teachers and pre-service teachers should possess.

For instance, Xu and Brown (2016) asserted that teachers should need to know “how different assessment methods can be related to the learning goals and specific content being learned, and what a variety of relevant assessment strategies are” (p.156), however did not underlined how to create assessment tools. Similarly, the Classroom Assessment Standards for PreK-12 Teachers (Klinger *et al.*, 2015), which were recently developed, do not place a strong emphasis on the creation of assessment tools. Instead, existing assessment standards tend to focus more on the effective use and interpretation of various assessment methods. On the other hand, educators might need not only to utilize established assessment tools but also to design diverse assessment tools that align with valid, reliable, and purposeful educational assessment principles. Existing tools may not possess all the intended properties or fully align with a specific learning goal or cognitive level. Thus, proposing this new dimension contributes by addressing this competency area within AL.

The current study further introduces “Competency in technology-based assessment” as a novel dimension. Similar to competency in developing assessment tools, none of the existing conceptual frameworks encompass a dimension related technology-based or digital assessment. However, technological advancements have created opportunities to enhance the efficiency and quality of assessments (Bennett, 2002). Besides, a recent systematic review study (Chen *et al.*, 2023), examining studies from 2018–2022, indicated that technology-based assessments demonstrated improved measurement accuracy, clarity, engagement, interaction, and enhanced teacher-parent communication. Thus, it can be asserted that it is inevitable to incorporate technology-based assessments within the context of assessment literacy. Assessment-literate teachers should effectively benefit from various types of technology-based assessments including computer-based tests, adaptive tests, simulations, games, and e-portfolios. Especially with the increasing advancements in artificial intelligence (AI), both teachers and pre-service teachers should be equipped with the necessary knowledge and skills for using and adapting AI-powered educational assessment tools improving the accuracy and efficiency of assessments, and enabling teachers to adapt their teaching strategies to meet the unique needs of each student (Gardner, O’Leary and Yuan, 2021; UNESCO,

2023). Therefore, this original component of the framework emphasizes the need to revolutionize the way education is delivered and assessed, in order to maximize the efficiency of teaching and learning processes via technology.

5.2. Discussion of RQ2: Reliability and Efficiency of the MST

The second purpose of the current research was to develop an MST for measuring assessment literacy. For this purpose, an MST was developed, and the second research question, “To what extent does the multistage test developed to measure assessment literacy provide reliable and efficient measurement across preservice teachers’ ability levels?” was investigated. Initially, a pilot test in a paper-and-pencil format was initially administered to 1223 participants as seven booklets based on a multiple matrix sampling (MMS). The quality of the items was evaluated through IRT-based analysis, considering their b parameters, item-fit statistics, ICCs, and IOC indices. This holistic approach was followed to create the item pool for the multistage test, rather than focusing solely on a single statistic or property of the items. As a result, a multistage test was designed and delivered to 106 participants through the Concerto platform.

Test results consisted of some anticipated findings as well as some undesirable findings. To illustrate, for pilot testing, the marginal reliability was calculated as 0.95 indicating excellent reliability, while the Cronbach’s alpha and the McDonald’s Omega coefficients calculated for each booklet fell below acceptable threshold (≥ 0.70). The internal consistency measures, such as alpha and omega, typically yield higher values when a test assesses a single construct reliably. However, if the test includes items measuring multiple constructs, this may result in lower reliability estimates (DeVellis, 2017). The formula for the alpha coefficient indicates that the more items in a test, the greater its reliability (DeVellis, 2017; Nunnally, 1978). Nevertheless, the fact that each dimension was represented by only 2 or 3 items in each booklet, totaling 20 items, may have contributed to the lower internal consistency coefficients observed. However, it might be significant to underline that the primary aim of pilot testing was calibrating item parameters required for the MST design. The limited number of

items per dimension in the booklets and the relatively small total item count in each booklet utilized during the pilot phase could have been the primary reasons the internal consistency coefficients did not reach the expected levels. On the other hand, the pilot phase successfully enabled pre-calibration of items. Furthermore, since the purpose of an IRT-based test is to measure the latent trait for all test-takers, considering an overall measure of the reliability of test scores across the latent distribution spectrum is crucial (Andersson and Xin, 2018). Notably, the calculated marginal reliability was found to be excellent (0.95). This IRT-based statistic shows that the test performs well across a wide range of ability levels, demonstrating consistent and reliable measurement precision regardless of the test-taker's proficiency.

However, the construct validity of the test structure could not be established based on the EFA and CFA results. Two of the booklets (Booklet 1 and Booklet 2) did not meet the significance threshold for conducting EFA according to Bartlett's test, and the KMO values for all booklets fell below the recommended 0.60 threshold. Consequently, the EFA yielded limited explained variance (29.95% - 37.38). Similarly, the CFA conducted for each booklet failed to generate a viable model. When the analysis was attempted in JASP (JASP Team, 2023), the model failed to converge due to the negative variance estimates, a phenomenon known as Heywood cases in the literature. Heywood cases occur when the communality value exceeds 1, which is theoretically implausible and indicates an issue with the model specification or the data quality (Farooq, 2022). This might suggest insufficient inter-item correlations or sample size limitations, both of which may hinder stable parameter estimation (Farooq, 2022).

These findings suggest that the test items do not effectively represent their intended factors, thereby limiting the pilot test results' construct validity. One potential reason for this might be the limited number of items (2-3) used to measure each of the 8 different dimensions. In the literature, there are various perspectives for the required number of items for each construct while conducting the factor analysis. For instance, Kline (2015) argues that a minimum of two indicators per factor is required to identify

CFA models with two or more factors. On the other hand, it is typically recommended to have a minimum of 3-5 items to measure each dimension (DeVellis, 2017; Nunnally and Bernstein, 1994). Therefore, the restricted item representation for each dimension contributed to the weak factor structure and low overall construct validity of pilot test results.

Despite the failure of the EFA and CFA analyses to yield the expected results, the booklet design, which incorporated questions from all dimensions, provided substantial benefits for test administration and the test's content validity. With an initial item pool of 105 items, administering the entire pool would have significantly increased the test duration, requiring four or five times more time. Additionally, the increased cognitive load on participants could have negatively impacted on their test performance. However, the booklet design approach ensured a well-balanced representation across the different dimensions, thereby supporting the test's content validity, as well as facilitating the test administration process.

Furthermore, the current study distinguished itself by following the Item Response Theory framework during data analysis, unlike many existing studies on assessment literacy instruments, which relied primarily on Classical Test Theory. While CTT-based analyses, such as computing Cronbach's alpha or calculating item difficulty based on the proportion of examinees answering correctly, were common, the present study benefited from both approaches. The IRT-based analysis provided additional evidence to support the findings, offering insights that CTT alone could not provide. For instance, the item information functions, which indicate how much information an item provides at different ability levels, are a type of analysis unique to IRT. Additionally, the study was able to create test information functions and calculate marginal reliability, which are other examples of IRT-based analyses. Moreover, although a 1-PL model was employed due to the limited sample size, the difficulty parameter derived from the IRT signifies the point at which examinees have a 50% probability of answering correctly, thereby rendering it sample-invariant (Baker and Kim, 2004).

The test relied on pre-calibrated test items that fit the 1-PL model, ensuring a wide range of b parameters across various difficulty levels.

Consequently, the results indicated that the MST aligned with the adaptive test paradigm. The test information functions generated for each module in the MST assessment demonstrated a well-balanced design. The routing module provided a solid foundation for guiding routing decisions. Furthermore, the modules in Stages 2 and 3 encompassed the ability range from approximately -2 to +2 on the theta scale. As a result, the overall structure ensured consistent information coverage across the full ability spectrum. Moreover, a decrease in participants' Standard Error of Measurement (SEM) across the stages of the MST demonstrated its effectiveness in adaptively refining ability estimates, as a lower SEM indicated higher measurement precision (van der Linden and Glas, 2010). Overall, the study successfully developed a multistage test for measuring assessment literacy despite several limitations.

5.3. The Promotion of Assessment Literacy

Cultivating assessment literacy is essential for pre-service and practicing educators to meet the complex demands of contemporary classrooms (DeLuca *et al.*, 2016). The identification of the scope of assessment literacy based on today's needs and opportunities (RQ1) highlights the complex nature of the construct as well as the significance of various assessment-related knowledge and skills. Amidst a considerable number of skills that teachers and preservice teachers should possess, assessment literacy should not be underestimated.

Teachers' conceptions of assessment tend to evolve based on acquiring knowledge about assessment (DeLuca and Lam, 2014). Thus, several factors affecting assessment literacy, such as the awareness of their role in assessment and the development of their identity as assessors, might be shaped with assessment literacy development (Xu and Brown, 2016). In that sense, it is crucial to prepare preservice teachers for future classroom environments requiring diverse competencies and to develop evidence-based

strategies to promote their assessment literacy. Designing assessment courses at the undergraduate level should be emphasized to capture the core dimensions of assessment literacy, with a special focus on emerging needs, such as the ability to conduct technology-based assessments. In Turkey, there is no standardized content for assessment courses, even though students must take at least one assessment course to graduate from the faculty of education, regardless of their department. Therefore, instructors should follow the current trends in the field of educational assessment, make instructional decisions that promote and enhance assessment literacy, and integrate crucial, essential content specified in the assessment framework (RQ1) into their teaching.

In addition to teacher education, other stakeholder-specific actions should be taken for promoting assessment literacy. Research revealed significant gaps in teacher self-efficacy regarding student learning evaluation, highlighting the need for enhanced professional development in assessment methodologies, supportive resources and dedicated time for assessment practices (Volante and Fazio, 2007). Hence, professional development programs designed in line with the current scope of assessment literacy can improve in-service teachers who feel deficiencies in this area, thereby enhancing their ability to use assessment as a tool for improving teaching and learning. On the other hand, as Brookhart (2011) underlines, a sustainable assessment literacy reform requires coordination between teacher preparation programs, professional learning communities, and coherent policy directives. The effectiveness of both teacher education and professional development programs ultimately depends on supportive educational policies. Therefore, for system-wide improvements, policymakers should share the concern for fostering assessment literacy among educators.

6. LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

This chapter examines the limitations uncovered in the research and provides several recommendations to further refine and progress the field of assessment literacy.

6.1. Limitations

While the current study marks substantial progress in measuring preservice teachers' assessment literacy through a multistage test that increases measurement precision, there is still considerable room to enhance the overall test development process, particularly in establishing the construct validity of the instrument. The study is grounded in the premise that assessment literacy is a multidimensional construct, and the test development process was intentionally designed to consider its diverse dimensions. However, due to the limited sample size and the representation of each dimension with only a few items in pilot testing, comprehensive construct validation was not feasible.

Another potential limitation is the possible violation of the unidimensionality assumption needed for the 1-PL IRT model. Despite evaluating unidimensionality using several methods, the results indicated multidimensionality, which could affect the interpretability and validity of the IRT analyses. Finally, the relatively small number of participants who were administered the MST may limit the generalizability of the findings, as the sample may not fully represent the broader target population. To address these limitations and strengthen the overall test development process, future research should focus on further refining and enhancing the design, validation, and administration of adaptive tests for measuring assessment literacy.

6.2. Suggestions for Further Research

The current study introduced a novel approach involving the creation of competency-based, measurable subdomains for assessment literacy. However, the individual dimensions were not examined separately. Future research should explore focused investigations of specific components within this assessment framework, particularly the newly proposed aspects related to technology-based assessment competencies and the ability to develop assessment tools. This in-depth examination of the distinct subdomain factors would provide valuable insights into the multifaceted nature of assessment literacy and how it can be most effectively measured and developed.

Additionally, while the structure of the assessment framework could not be empirically tested in the current study, validity evidence may be obtained through further research using various methodologies. For instance, a concurrent validation study could investigate the correlation between the theta scores from the MST and another measure of assessment competencies, such as final exam scores in an assessment course. This would reveal the extent to which performance on the MST aligns with performance on other relevant assessments, thereby offering insights into the practical utility and real-world applications of the test scores. Furthermore, to enhance the validity, the pilot test in paper-and-pencil format as well as the multistage test could be re-administered with a larger sample size.

Moreover, performance level descriptors could be developed to provide meaningful interpretations of the MST scores, as these descriptors have the potential to reflect a test-taker's knowledge and abilities within specific score ranges. By establishing clear performance level descriptors, the assessment results can be communicated and understood more effectively, enabling stakeholders to gain a deeper understanding of the test-taker's assessment literacy competencies. In addition to such descriptors, a guideline for administration, scoring, and interpretation of test scores might be developed, as Downing and Haladyna (2006) suggest. This phase of test development could not be implemented within the current study due to time constraints.

REFERENCES

- Abell, S. K., and M.A. Siegel, *The Professional Knowledge Base of Science Teaching*, Springer, Dordrecht, 2011.
- Alzhrany, A., *An Inquiry into Teacher Assessment Literacy: The Development of an Assessment Literacy Framework and Instrument to Measure the Assessment Literacy of Teachers in Saudi Arabia*, Ph.D. Thesis, The University of Sydney, 2022.
- Andersson, B. and Xin, T., 2018, “Large Sample Confidence Intervals for Item Response Theory Reliability Coefficients”, *Educational and Psychological Measurement*, Vol. 78, No. 1, pp. 32-45.
- Aybek, E.C., Arıkan, S. and Ertas, G., 2024, “A Practical Guide to Item Bank Calibration with Multiple Matrix Sampling”, *International Journal of Assessment Tools in Education*, Vol. 11, No. 4, pp. 647–659.
- Azrak, Y. and Yalçınkaya, E., 2019, “Sosyal Bilgiler Öğretmen Adaylarının Ölçme-Değerlendirme Okuryazarlık Düzeylerinin Çeşitli Değişkenler Açısından İncelenmesi”, *Anadolu Journal of Educational Sciences International*, Vol. 9, No. 1, pp. 27–55.
- Baker, F., *The Basics of Item Response Theory*, Second Edition, ERIC Clearinghouse on Assessment and Evaluation, Washington, 2001.
- Baker, F.B. and Kim, S.H., 2004, *Item Response Theory: Parameter Estimation Techniques*, Chain Reaction Cycles.
- Bartlett, M.S., 1950, “Tests of Significance in Factor Analysis”, *British Journal of Psychology*, Vol. 3, No.1, pp. 77–85.

- Bennett, R.E., 2002, “Inexorable and Inevitable: The Continuing Story of Technology and Assessment”, *Computer-Based Testing and the Internet*, Vol. 1, No. 1, pp. 201–217.
- Brookhart, S., 2011, “Educational Assessment Knowledge and Skills for Teachers”, *Educational Measurement: Issues and Practice*, Vol. 30, No. 1, pp. 3–12.
- Brown, D. and Clair, D., “Integrated R-transportation”, *Paper Presented at the 14th International Conference of Transportation Logistics*, Istanbul, 2008.
- Bütüner, S.Ö., Yiğit, N. and Çimer, S.O., 2010, “Ölçme Değerlendirme Okuryazarlığı Envanterinin Türkçeye Uyarlanması”, *E-Journal of New World Sciences Academy*, Vol. 5, No. 3, pp. 792–809.
- Campbell, C., Murphy, J.A. and Holt, J.K., “Psychometric Analysis of an Assessment Literacy Instrument: Applicability to Pre-Service Teachers”, *Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association*, Columbus, 2002.
- Cattell, R.B., 1966, “The Scree Test for the Number of Factors”, *Multivariate Behavioral Research*, Vol. 1, No. 2, pp. 245–276.
- Chalmers, R.P., 2012, “Mirt: A Multidimensional Item Response Theory Package for the R Environment”, *Journal of Statistical Software*, Vol. 48, No. 6, pp. 1–29.
- Chan, C.K.Y. and Luk, L.Y.Y., 2022, “A Four-Dimensional Framework for Teacher Assessment Literacy in Holistic Competencies”, *Assessment and Evaluation in Higher Education*, Vol. 47, No. 5, pp. 755–769.
- Chen, D., Jeng, A., Sun, S. and Kaptur, B., 2023, “Use of Technology-Based Assessments: A Systematic Review Covering Over 30 Countries”, *Assessment in Educa-*

tion: Principles, Policy and Practice, Vol. 30, No. 5–6, pp. 396–428.

Cheng, Y., Yuan, K.-H. and Liu, C., 2012, “Comparison of Reliability Measures Under Factor Analysis and Item Response Theory”, *Educational and Psychological Measurement*, Vol. 72, No. 1, pp. 52–67.

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.

Creswell, J.W., *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Fourth Edition, Sage, California, 2014.

Crocker, L. and Algina, J., *Introduction to Classical and Modern Test Theory*, Wadsworth, Belmont, California, 2006.

Cronbach, L.J., 1951, “Coefficient Alpha and the Internal Structure of Tests”, *Psychometrika*, Vol. 16, No. 3, pp. 297–334.

DeLuca, C. and Lam, C., 2014, “Preparing Teachers for Assessment Within Diverse Classrooms: An Analysis of Teacher Candidates’ Readiness”, *Teacher Education Quarterly*, Vol. 41, No. 3, pp. 3–24.

DeLuca, C., LaPointe-McEwan, D. and Luhanga, U., 2016, “Teacher Assessment Literacy: A Review of International Standards and Measures”, *Educational Assessment, Evaluation and Accountability*, Vol. 28, No. 1, pp. 251–272.

Demirdöğen, B. and Korkut, H.M., 2021, “Does Teacher Education Matter? Comparison of Education and Science Major Teachers’ Assessment Literacy”, *Eğitimde Nitel Araştırmalar Dergisi*, Vol. 26, No. 1, pp. 23–52.

DeVellis, R.F., 2017, *Scale Development: Theory and Applications*, Fourth Edition, Sage.

Dinan-Thomas, M. and Penney, D., 2015, "Assessment Literacy in Primary Physical Education", *European Physical Education Review*, Vol. 21, No. 4, pp. 485–503.

Downing, S.M. and Haladyna, T.M. (Eds.), *Handbook of Test Development*, Lawrence Erlbaum Associates Publishers, New York, 2006.

Dunn, T.J., Baguley, T. and Brunsden, V., 2014, "From Alpha to Omega: A Practical Solution to The Pervasive Problem of Internal Consistency Estimation", *British Journal of Psychology*, Vol. 105, No. 3, pp. 399–412.

Embretson, S.E. and Reise, S.P., *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates Publishers, New York, 2000.

Erdost, A., *Türk İngilizce Okutmanlarının Yabancı Dilde Ölçme ve Değerlendirme Okuryazarlığı: Çoklu Bir Durum Araştırması*, Ph.D. Thesis, Atatürk University, 2018.

Ergül, A.Ö. and Çetin, S., 2021, "Measurement and Assessment Literacy Levels of Teachers in Terms of Some Variables", *Adıyaman University Journal of Educational Sciences*, Vol. 11, No. 1, pp. 26–35.

Erol, T. and Tunagür, M., 2024, "Türkçe Öğretmenlerinin Ölçme Değerlendirme Okuryazarlık Düzeylerinin İncelenmesi", *İçtimaiyat (Türk Eğitim Sisteminde Değişimler ve Yeni Eğilimler)*, Vol. 8, No. 2, pp. 149–165.

Farooq, R., 2022, "Heywood Cases: Possible Causes and Solutions", *International Journal of Data Analysis Techniques and Strategies*, Vol. 14, No. 1, pp. 79–96.

- Gardner, J., O’Leary, M. and Yuan, L., 2021, “Artificial Intelligence in Educational Assessment: ‘Breakthrough? Or Buncombe and Ballyhoo”, *Journal of Computer Assisted Learning*, Vol. 37, No. 5, pp. 1207–1216.
- George, D. and Mallery, P., *SPSS for Windows Step by Step: A Simple Guide and Reference, Fourth Edition*, Allyn and Bacon, Boston, 2003.
- Girgla, A., Good, L., Krstic, S., McGinley, B., Richardson, S., Sneider-Gregory, S. and Star, J., 2021, *Developing a Teachers’ Assessment Literacy and Design Competence Framework*, Australian Council for Educational Research.
- Gonzalez, E. and Rutkowski, L., 2010, “Principles of Multiple Matrix Booklet Designs and Parameter Recovery in Large-Scale Assessments”, *IEA-ETS Research Institute Monograph*, Vol. 1, No. 1, pp. 125–156.
- Gotch, C.M. and French, B.F., 2014, “A Systematic Review of Assessment Literacy Measures”, *Educational Measurement: Issues and Practice*, Vol. 33, No. 1, pp. 14–18.
- Green, S.B., Levy, R., Thompson, M.S., Lu, M. and Lo, W., 2012, “A Proposed Solution to The Problem with Using Completely Random Data to Assess the Number of Factors with Parallel Analysis”, *Educational and Psychological Measurement*, Vol. 72, No. 3, pp. 357–374.
- Gül, E., *Preservice Elementary Teachers’ Assessment Literacy and Attitude Towards Assessment*, M.S. Thesis, Firat University, 2011.
- Haladyna, T.M., Rodriguez, M.C. and Downing, S.M., *Developing and Validating Test Items*, Routledge, New York, 2013.
- Hambleton, R.K. and Swaminathan, H., 1985, *Item Response Theory: Principles and*

Applications, Springer Science and Business Media.

- Hambleton, R.K. and Zenisky, A.L., 2013, “Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design”, *American Psychological Association*, Vol. 3. No. 1, pp. 479–494.
- Hay, P. and Penney, D., 2009, “Proposing Conditions for Assessment Efficacy in Physical Education”, *European Physical Education Review*, Vol. 15, No. 3, pp. 389–405.
- Hayes, A.F. and Coutts, J.J., 2020, “Use Omega Rather Than Cronbach’s Alpha for Estimating Reliability. But”, *Communication Methods and Measures*, Vol. 14, No. 1, pp. 1–24.
- Hendrickson, A., 2007, “An NCME Instructional Module on Multistage Testing”, *Educational Measurement: Issues and Practice*, Vol. 26, No. 2, pp. 44–52.
- Horn, J.L., 1965, “A Rationale and Test for the Number of Factors in Factor Analysis”, *Psychometrika*, Vol. 30, No. 2, pp. 179–185.
- Hu, L.T. and Bentler, P.M., 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.
- John, O.P. and Benet-Martínez, V., 2000, “Measurement: Reliability, Construct Validation, And Scale Construction”, *Cambridge University Press*, VOL. 1, No. 1, pp. 339–369.
- Kaiser, H.F., 1960, “The Application of Electronic Computers to Factor Analysis “, *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 141–151.
- Karaman, P. and Şahin, Ç., 2014, “Öğretmen Adaylarının Ölçme Değerlendirme Okur

- yazarlıklarının Belirlenmesi”, *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, Vol. 15, No. 2, pp. 175–189.
- Kauffman, D., Johnson, S.M., Kardos, S.M., Liu, E. and Peske, H.G., 2022, “Lost at Sea: New Teachers’ Experiences with Curriculum and Assessment”, *Teachers College Record*, Vol. 104, No. 2, pp. 273–300.
- Kim, S., 2012, “A Note on the Reliability Coefficients for Item Response Model-Based Ability Estimates”, *Psychometrika*, Vol. 77, No. 1, pp. 153–162.
- Kline, R.B., 2015, *Principles and Practice of Structural Equation Modeling*, Fourth Edition, Guilford Press.
- Klinger, D.A., McDivitt, P.R., Howard, B.B., Munoz, M.A., Rogers, W.T. and Wylie, E.C., 2015, *The Classroom Assessment Standards for PreK-12 Teachers*, Kindle Direct Press.
- Koç, A. and Bulut, S., 2020, “Fen Bilgisi Öğretmen Adaylarının Ölçme Değerlendirme Okuryazarlık Düzeylerinin Çeşitli Değişkenler Açısından İncelenmesi”, *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, Vol. 40, No. 3, pp. 877–904.
- Koh, K., Burke, L.E.C., Luke, A., Gong, W. and Tan, C., 2017, “Developing the Assessment Literacy of Teachers in Chinese Language Classrooms: A Focus on Assessment Task Design”, *Language Teaching Research*, Vol. 22, No. 3, pp. 264–288.
- Lawson, H.A., 1986, “Occupational Socialization and the Design of Teacher Education Programs”, *Journal of Teaching in Physical Education*, Vol. 5, No. 2, pp. 107–116.
- Lord, F.M., 1971, “A Theoretical Study of Two-Stage Testing”, *Psychometrika*, Vol. 36, No. 1, pp. 227–242.

- Lord, F.M., 1980, *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates.
- Lord, F.M., 1986, “Maximum Likelihood and Bayesian Parameter Estimation In Item Response Theory”, *Journal of Educational Measurement*, Vol. 23, No. 2, pp. 157–162.
- Luecht, R.M. and Sireci, S.G., 2012, *A Review of Models for Computer-Based Testing*, Research Report No. RR-2011-12, The College Board, New York.
- Luecht, R.M., 1998, “Computer-Assisted Test Assembly Using Optimization Heuristics”, *Applied Psychological Measurement*, Vol. 22, No. 3, pp. 224–236.
- Luecht, R.M., 2000, “Implementing the Computer-Adaptive Sequential Testing (CAST) Framework to Mass Produce High Quality Computer-Adaptive and Mastery Tests”, *paper presented at the Annual Meeting of the National Council on Measurement in Education*, New Orleans, Los Angeles.
- Lumadi, M.W., 2013, “Challenges Besetting Teachers in Classroom Assessment: An Exploratory Perspective”, *Journal of Social Sciences*, Vol. 34, No. 3, pp. 211–221.
- Magis, D., Yan, D. and Von Davier, A.A., 2017, *Computerized Adaptive and Multistage Testing with R*, Springer, Switzerland.
- McMillan, J.H., 2001, “Secondary Teachers’ Classroom Assessment and Grading Practices”, *Educational Measurement: Issues and Practice*, Vol. 20, No. 1, pp. 20–32.
- McMillan, J.H., 2013, “Why We Need Research on Classroom Assessment”, *Why We Need Research on Classroom Assessment*, SAGE Publications, Inc., Vol. 1, No. 1, pp. 2–16.

- Mertler, C.A. and Campbell, C., 2005, “Measuring Teachers’ Knowledge and Application of Classroom Assessment Concepts: Development of the Assessment Literacy Inventory”, *The annual meeting of the American Educational Research Association*, Montreal, Canada.
- Mertler, C.A., 2003, “Pre-Service Versus In-Service Teachers’ Assessment Literacy: Does Classroom Experience Make a Difference?”, *The annual meeting of the Mid-Western Educational Research Association*, Columbus, Ohio.
- Messick, S., 1989, “Validity”, in R.L. Linn (Ed.), *Educational Measurement*, Third Edition, American Council on Education and Macmillan, New York.
- Mullis, I.V.S., Martin, M.O. and von Davier, M., 2021, *TIMSS 2023 Assessment Frameworks*, International Association for the Evaluation of Educational Achievement, Boston, Vol. 1, No. 1, pp. 15–16.
- Nunnally, J.C. and Bernstein, I.H., 1994, *Psychometric Theory*, Third Edition, McGraw-Hill, New York.
- Nunnally, J.C., 1978, *Psychometric Theory*, Second Edition, McGraw-Hill, New York.
- Ogan-Bekiroğlu, F. and Suzuk, E., 2014, “Pre-Service Teachers’ Assessment Literacy and Its Implementation into Practice”, *The Curriculum Journal*, Vol. 25, No. 3, pp. 344–371.
- Orlando, M. and Thissen, D., 2000, “Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models”, *Applied Psychological Measurement*, Vol. 24, No. 1, pp. 50–64.
- Padilla, M.A. and Divers, J., 2013, “Bootstrap Interval Estimation of Reliability Via Coefficient Omega”, *Journal of Modern Applied Statistical Methods*, Vol. 12, No. 1,

pp. 78–89.

Pastore, S. and Andrade, H.L., 2019, “Teacher Assessment Literacy: A Three-Dimensional Model”, *Teaching and Teacher Education*, Vol. 84, No. 1, pp. 128–138.

Pastore, S., 2023, “Teacher Assessment Literacy: A Systematic Review”, *Frontiers in Education*, Vol. 8, No. 1, pp. 1217167–1217198.

Phelps, R.P., 2015, “Synergies for Better Learning: An International Perspective on Evaluation and Assessment”, SSRN Electronic Journal.

Plake, B.S., Impara, J.C. and Fager, J.J., 1993, “Assessment Competencies of Teachers: A National Survey”, *Educational Measurement: Issues and Practice*, Vol. 12, No. 4, pp. 10–12.

Popham, W.J., 1981, *Modern Educational Measurement*, Prentice-Hall, Englewood Cliffs, New Jersey.

Popham, W.J., 2009, “Assessment Literacy for Teachers: Faddish or Fundamental”, *Theory into Practice*, Vol. 48, No. 1, pp. 4–11.

R Core Team, 2023, “R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna”, <https://www.R-project.org/>, accessed on January 4, 2024.

Rasch, G., 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen.

Revelle, W., “psych: Procedures for Psychological, Psychometric, and Personality Research (Version 2.3.6)”, 2023, <https://CRAN.R-project.org/package=psych>, accessed on June 4, 2024.

- Rovinelli, R.J. and Hambleton, R.K., 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.
- Sari, H.I., Yahsi-Sari, H. and Huggins-Manley, A.C., 2016, "Computer Adaptive Multi-stage Testing: Practical Issues, Challenges and Principles", *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, Vol. 7, No. 2, pp. 388–407.
- Shepard, L.A., 2000, "The Role of Assessment in a Learning Culture", *Educational Researcher*, Vol. 29, No. 7, pp. 4–14.
- Shoemaker, D.M., 1973, *Principles and Procedures of Multiple Matrix Sampling*, Ballinger Publishing Company, Cambridge.
- Smits, N., Zitman, F.G., Cuijpers, P., den Hollander-Gijsman, M.E. and Carlier, I.V., 2012, "A Proof of Principle for Using Adaptive Testing in Routine Outcome Monitoring: The Efficiency of the Mood and Anxiety Symptoms Questionnaire–Anhedonic Depression CAT", *BMC Medical Research Methodology*, Vol. 12, No. 1, pp. 8–9.
- Starck, J.R., Richards, K.A.R. and O'Neil, K., 2018, "A Conceptual Framework for Assessment Literacy: Opportunities for Physical Education Teacher Education", *Quest*, Vol. 70, No. 4, pp. 519–535.
- Stiggins, R.J., 1991, "Assessment Literacy", *Phi Delta Kappan*, Vol. 72, No. 7, pp. 534–539.
- Stiggins, R.J., 1995, "Assessment Literacy for The 21st Century", *Phi Delta Kappan*, Vol. 77, No. 3, pp. 238–245.
- Tünkler, V., 2019, "Sosyal Bilgiler Öğretmen Adaylarının Değerlendirme Okuryazarlık Düzeyi ile Değerlendirme Öz-Yeterlik Algılarının İncelenmesi", *Pamukkale Üniversitesi*

Sosyal Bilimler Enstitüsü Dergisi, Vol. 37, No. 1, pp. 319–332.

Turner, R.C. and Carlson, L., 2003, “Indexes of Item-Objective Congruence for Multi-dimensional Items”, *International Journal of Testing*, Vol. 3, No. 2, pp. 163–171.

Ullman, J.B., 2001 “Structural Equation Modeling”, in B.G. Tabachnick and L.S. Fidell (Eds.), *Using Multivariate Statistics*, Pearson Education, 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.

UNESCO, Miao, F. and Holmes, W., 2023, *UNESCO Guidance for Generative AI in Education and Research*, United Nations Educational, Scientific and Cultural Organisation.

Van Der Linden, W.J. and Glas, C.A., 2010, *Elements of Adaptive Testing*, Springer, New York.

Volante, L. and Fazio, X., 2007, “Exploring Teacher Candidates’ Assessment Literacy: Implications for Teacher Education Reform and Professional Development”, *Canadian Journal of Education/Revue Canadienne de l’éducation*, Vol. 30, No. 3, pp. 749–770.

Wainer, H., Lewis, C., Kaplan, B. and Braswell, J., 1990, *An Adaptive Algebra Test: A Testlet-Based, Hierarchically Structured Test with Validity-Based Scoring*, Technical Report No. 90-92, Educational Testing Service.

Weissman, A., Belov, D. and Armstrong, R.D., 2007, *Information-Based Versus Number-Correct Routing in Multistage Classification Tests*, LSAC Research Report No. 07-05, Law School Admission Council.

- Willis, J., Adie, L. and Klenowski, V., 2013, "Conceptualising Teachers' Assessment Literacies in an Era of Curriculum and Assessment Reform", *The Australian Educational Researcher*, Vol. 40, No. 2, pp. 241–256.
- Xu, Y. and Brown, G., 2016, "Teacher Assessment Literacy in Practice: A Reconceptualization", *Teaching and Teacher Education*, Vol. 58, No. 1, pp. 149–162.
- Yan, Z. and Pastore, S., 2022, "Are Teachers Literate in Formative Assessment? The Development and Validation of the Teacher Formative Assessment Literacy Scale", *Studies in Educational Evaluation*, Vol. 74, No. 1, pp. 101–183.
- Yaşar, M.D., 2020, "Chemistry Teachers' Assessment Literacy in Turkey", *Journal of Baltic Science Education*, Vol. 19, No. 1, pp. 76–90.
- 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, No. 1, pp. 125–145.
- Zenisky, A., Hambleton, R.K. and Luecht, R.M., 2010, "Multistage Testing: Issues, Designs, And Research", in W.J. van der Linden and C.A.W. Glas (Eds.), *Elements of Adaptive Testing*, Springer, Vol. 1, No. 1, pp. 355–372.
- Zenisky, A.L., 2004, *Evaluating the Effects of Several Multi-Stage Testing Design Variables on Selected Psychometric Outcomes for Certification and Licensure Assessment*, Ph.D. Thesis, University of Massachusetts Amherst.
- Zheng, Y. and Chang, H.-H., 2015, "On-The-Fly Assembled Multistage Adaptive Testing", *Applied Psychological Measurement*, Vol. 39, No. 2, pp. 104–118.

APPENDIX A: IOC INDICES

IOC indices for each item in the item pool for pilot phase are given in Appendix A.

Table A.1. IOC indices.

Item ID	IOC Index
K3B1S1K	10.00
K3B2S3K	10.00
K2B8S2K	00.93
K3B7S1K	00.89
K2B5S1K	00.93
K1B2S1O	10.00
K1B1S1O	00.43
K1B3S1O	-00.14
K1B4S1O	10.00
K1B5S1O	10.00
K2B6S1Z	-00.18
K2B7S1Z	10.00
K2B8S1Z	-00.14
K3B4S1Z	00.96
K3B3S2Z	00.96
K1B8S2K	10.00
K2B6S2K	00.39
K1B1S1K	00.32
K3B2S1K	00.96
K3B4S1K	00.93
K3B3S1K	00.75

Table A.1 IOC indices. (cont.)

Item ID	IOC Index
K3B5S1K	-00.18
K1B6S1O	00.96
K1B7S1O	00.89
K1B8S1O	10.00
K2B1S1O	-00.14
K2B2S1O	00.93
K3B6S1Z	-00.14
K3B7S1Z	10.00
K2B8S2Z	-00.29
K3B8S1Z	00.43
K1B5S1Z	00.93
K2B2S1K	00.32
K2B5S2K	-00.14
K1B5S1K	00.43
K1B2S1K	10.00
K1B3S1K	-00.21
K2B8S3K	-00.14
K1B4S1K	10.00
K2B3S1O	00.36
K2B4S1O	10.00
K2B5S1O	00.43
K2B6S1O	10.00
K2B7S1O	10.00
K1B6S1Z	10.00
K1B7S1Z	10.00
K1B8S2Z	00.93
K1B2S1Z	10.00
K2B1S1Z	00.43

Table A.1 IOC indices. (cont.)

Item ID	IOC Index
K3B3S2K	00.93
K2B8S2O	00.43
K1B6S1K	10.00
K1B7S1K	10.00
K2B3S1K	00.39
K3B2S2K	00.75
K3B6S1K	00.93
K3B1S1O	00.93
K3B2S1O	10.00
K2B8S1O	00.11
K3B3S1O	00.96
K3B4S1O	10.00
K2B2S1Z	00.64
K2B3S1Z	00.43
K2B4S1Z	10.00
K1B1S1Z	10.00
K3B5S1Z	00.89
K3B5S1O	10.00
K3B7S1O	10.00
K3B6S1O	00.39
K2B1S1K	-00.14
K3B1S2K	00.43
K2B7S1K	00.93
K2B8S1K	00.36
K2B6S1K	00.36
K2B5S2O	00.89
K2B6S2O	00.36
K2B7S2O	00.93
K1B8S2O	00.50

Table A.1 IOC indices. (cont.)

Item ID	IOC Index
K3B8S1O	00.36
K1B3S1Z	00.93
K1B4S1Z	10.00
K2B8S3Z	-00.14
K2B5S1Z	00.29
K3B2S1Z	10.00
K2B8S3O	-00.14
K2B7S2Z	00.89
K3B2S3Z	10.00
K3B4S2K	10.00
K2B7S2K	10.00
K1B8S1K	00.96
K3B8S1K	00.43
K2B4S1K	10.00
K3B1S2O	10.00
K3B2S2O	10.00
K3B2S3O	10.00
K3B3S2O	10.00
K3B4S2O	10.00
K3B1S1Z	10.00
K3B2S2Z	10.00
K2B5S2Z	10.00
K2B6S2Z	00.39
K3B3S1Z	10.00
K1B8S1Z	10.00
K3B1S2Z	10.00
K3B4S2Z	10.00

APPENDIX B: EFA RESULTS

Exploratory Factor Analysis (EFA) results for each booklet administered in pilot phase are given in Appendix B.

Table B.1. KMO and Bartlett's test for booklet 1.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy 0.		0.497
Bartlett's Test of Sphericity	Approx 0. Chi-Square	199.586
	df	190
	Sig 0.	0.302

Table B.2. Communalities for booklet 1.

	Initial	Extraction
K3B1S1K	0.107	0.216
K3B2S3K	0.081	0.164
K2B8S2K	0.114	0.332
K3B7S1K	0.081	0.219
K2B5S1K	0.101	0.395
K1B2S1O	0.130	0.312
K1B1S1O	0.141	0.639
K1B3S1O	0.154	0.342
K1B4S1O	0.105	0.277
K1B5S1O	0.087	0.304
K1B6S1O	0.072	0.127
K1B7S1O	0.119	0.272
K1B8S1O	0.116	0.397
K2B1S1O	0.214	0.604
K2B2S1O	0.088	0.218
K2B6S1Z	0.113	0.199
K2B7S1Z	0.110	0.360

Table B.2 Communalities for booklet 1. (cont.)

K2B8S1Z	0.146	0.509
K3B4S1Z	0.109	0.297
K3B3S2Z	0.104	0.209

Table B.3. Variance explained and factor eigenvalues for booklet 1.

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.765	8.826	8.826	1.178	5.892	5.892
2	1.550	7.750	16.576	0.893	4.464	10.356
3	1.388	6.940	23.516	0.751	3.756	14.112
4	1.364	6.820	30.336	0.684	3.418	17.531
5	1.324	6.621	36.958	0.659	3.297	20.827
6	1.201	6.005	42.963	0.573	2.863	23.691
7	1.168	5.838	48.801	0.543	2.716	26.407
8	1.121	5.606	54.407	0.416	2.082	28.488
9	1.067	5.336	59.743	0.376	1.880	30.368
10	1.011	5.056	64.799	0.318	1.591	31.959
11	0.910	4.551	69.349			
12	0.851	4.256	73.605			
13	0.786	3.930	77.535			
14	0.776	3.878	81.413			
15	0.737	3.684	85.097			
16	0.680	3.402	88.499			
17	0.665	3.323	91.822			
18	0.571	2.856	94.677			
19	0.567	2.835	97.512			
20	0.498	2.488	100.000			

Table B.4. Factor matrix for booklet 1.

	Factor									
	1	2	3	4	5	6	7	8	9	10
K2B1S1O	0.612			-0.416						
K1B4S1O	0.317									
K1B7S1O										
K1B3S1O		0.444								
K2B8S1Z	-0.313	0.412					0.315			
K2B6S1Z		0.325								
K1B2S1O		0.315								
K1B1S1O	0.394		0.536			0.328				
K2B8S2K			-0.317							
K1B6S1O										
K3B3S2Z										
K3B4S1Z					-0.330					
K2B2S1O					0.309					
K3B1S1K					0.308					
K1B5S1O					0.305					
K3B2S3K										
K1B8S1O						0.344				
K2B5S1K								0.313		
K3B7S1K										
K2B7S1Z									0.343	

Table B.5. Pattern matrix for booklet 1.

Pattern Matrix	
Rotation failed to converge in 25 iterations. (Convergence = .001).	

Table B.6. KMO and Bartlett's test for booklet 2.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.507
Bartlett's Test of Sphericity	Approx. Chi-Square	258.738
	df	190
	Sig.	<0.001

Table B.7. Communalities for booklet 2.

	Initial	Extraction
K1B1S1K	0.235	0.611
K3B2S1K	0.161	0.194
K3B4S1K	0.118	0.166
K3B3S1K	0.169	0.312
K3B5S1K	0.170	0.460
K1B6S1O	0.127	0.547
K1B7S1O	0.172	0.528
K1B8S1O	0.142	0.228
K2B1S1O	0.121	0.361
K2B2S1O	0.147	0.411
K2B3S1O	0.177	0.326
K2B4S1O	0.148	0.276
K2B5S1O	0.153	0.206
K2B6S1O	0.084	0.136
K2B7S1O	0.191	0.759
K3B6S1Z	0.208	0.447
K3B7S1Z	0.180	0.348
K2B8S2Z	0.168	0.287
K3B8S1Z	0.085	0.095
K1B5S1Z	0.174	0.253

Table B.8. Variance explained and factor eigenvalues for booklet 2.

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumu- lative %	Total	% of Variance	Cumu- lative %	Total
Factor							
1	2.047	10.235	10.235	1.415	7.075	7.075	0.989
2	1.684	8.421	18.656	1.112	5.559	12.634	0.966
3	1.449	7.245	25.901	0.910	4.552	17.186	0.937
4	1.432	7.159	33.060	0.799	3.996	21.182	0.723
5	1.273	6.366	39.426	0.716	3.578	24.760	0.776
6	1.214	6.070	45.495	0.578	2.890	27.650	0.728
7	1.189	5.946	51.441	0.547	2.734	30.384	0.892
8	1.068	5.338	56.779	0.462	2.309	32.694	0.833
9	1.062	5.308	62.087	0.410	2.052	34.746	0.688
10	0.963	4.816	66.903				
11	0.880	4.400	71.302				
12	0.857	4.287	75.589				
13	0.811	4.054	79.643				
14	0.712	3.562	83.205				
15	0.676	3.382	86.587				
16	0.631	3.155	89.742				
17	0.593	2.965	92.708				
18	0.555	2.774	95.482				
19	0.460	2.301	97.783				
20	0.443	2.217	100.000				

Table B.9. Factor matrix for booklet 2.

	Factor								
	1	2	3	4	5	6	7	8	9
K3B6S1Z	0.514								
K2B2S1O	0.373						-0.301		
K2B8S2Z	0.363					-0.310			
K3B2S1K	0.330								
K2B3S1O	0.324								
K2B4S1O									
K1B1S1K		0.678							
K2B5S1O		0.336							
K3B8S1Z									
K2B7S1O			0.788						
K2B1S1O				0.343					
K1B8S1O				0.316					
K3B3S1K					-0.433				
K3B7S1Z	0.342				-0.378				
K1B5S1Z					0.366				
K1B6S1O				0.433		0.448			
K3B5S1K	.321						0.391		
K3B4S1K							-0.336		
K1B7S1O	0.340	0.380						-0.422	
K2B6S1O									

Table B.10. KMO and Bartlett's test for booklet 3.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.553
Bartlett's Test of Sphericity	Approx. Chi-Square	253.956
	df	190
	Sig.	0.001

Table B.11. Communalities for booklet 3.

	Initial	Extraction
K1B5S1K	0.161	0.239
K1B2S1K	0.166	0.332
K1B3S1K	0.141	0.161
K2B8S3K	0.102	0.145
K1B4S1K	0.207	0.290
K2B3S1O	0.217	0.286
K2B4S1O	0.150	0.342
K2B5S1O	0.091	0.113
K2B6S1O	0.092	0.147
K2B7S1O	0.089	0.126
K3B1S1O	0.098	0.185
K3B2S1O	0.144	0.398
K2B8S1O	0.219	0.443
K3B3S1O	0.146	0.477
K3B4S1O	0.166	0.476
K1B6S1Z	0.271	0.618
K1B7S1Z	0.160	0.235
K1B8S2Z	0.155	0.697
K1B2S1Z	0.169	0.380
K2B1S1Z	0.153	0.212

Table B.12. Variance explained and factor eigenvalues for booklet 3.

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.225	11.126	11.126	1.610	8.051	8.051	1.173
2	1.711	8.556	19.682	1.040	5.198	13.249	0.969
3	1.455	7.275	26.958	0.864	4.322	17.571	0.893
4	1.380	6.898	33.856	0.769	3.845	21.416	0.812
5	1.238	6.189	40.044	0.634	3.168	24.585	0.642
6	1.206	6.032	46.076	0.510	2.549	27.133	0.618
7	1.143	5.716	51.792	0.460	2.302	29.436	0.796
8	1.047	5.236	57.028	0.417	2.083	31.519	0.940
9	0.981	4.906	61.934				
10	0.935	4.674	66.607				
11	0.914	4.570	71.177				
12	0.894	4.468	75.645				
13	0.760	3.802	79.446				
14	0.728	3.640	83.087				
15	0.671	3.357	86.444				
16	0.633	3.166	89.610				
17	0.588	2.939	92.550				
18	0.540	2.699	95.248				
19	0.486	2.428	97.676				
20	0.465	2.324	100.000				

Table B.13. Factor matrix for booklet 3.

	Factor							
	11	22	33	44	55	66	77	88
K1B6S1Z	0.614			-0.386				
K2B8S1O	0.507				0.345			
K2B3S1O	0.458							
K1B4S1K	0.366			-0.328				
K2B4S1O	0.360							
K1B3S1K								
K3B3S1O		0.435				0.332	-0.303	
K1B2S1K		0.385						
K1B2S1Z		-0.332						
K2B1S1Z								
K1B7S1Z								
K1B5S1K								
K2B8S3K								
K1B8S2Z			0.698					
K3B4S1O		-0.326		0.362				
K3B1S1O								
K2B7S1O								
K3B2S1O						-0.331		
K2B5S1O								
K2B6S1O								0.317

Table B.14. Pattern matrix for booklet 3.

	Factor							
	1	2	3	4	5	6	7	8
K1B6S1Z	0.707							
K1B4S1K	0.532							

Table B.14 Pattern matrix for booklet 3.(cont.)

	Factor							
K1B3S1K	0.349							
K2B5S1O								
K1B2S1K		0.558						
K2B4S1O		0.453			0.302			
K2B8S3K		0.331						
K2B1S1Z								
K1B8S2Z			0.847					
K1B5S1K								
K3B4S1O				0.680				
K1B7S1Z				0.371				
K3B1S1O					-0.420			
K2B7S1O					0.321			
K2B6S1O								
K3B2S1O						-0.587		
K3B3S1O							-0.670	
K1B2S1Z							0.403	
K2B8S1O								-0.594
K2B3S1O								-0.388

Table B.15. KMO and Bartlett's test for booklet 4.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.550
Bartlett's Test of Sphericity	Approx. Chi-Square	222.627
	df	190
	Sig.	0.053

Table B.16. Communalities for booklet 4.

<i>Communalities</i>		
	Initial	Extraction
K1B6S1K	0.080	0.140
K1B7S1K	0.153	0.402
K2B3S1K	0.210	0.475
K3B2S2K	0.153	0.244
K3B6S1K	0.172	0.302
K3B1S1O	0.105	0.307
K3B2S1O	0.156	0.458
K2B8S1O	0.133	0.223
K3B3S1O	0.097	0.177
K3B4S1O	0.143	0.407
K2B5S2O	0.131	0.250
K2B6S2O	0.125	0.260
K2B7S2O	0.113	0.392
K1B8S2O	0.135	0.331
K3B8S1O	0.154	0.683
K2B2S1Z	0.084	0.260
K2B3S1Z	0.097	0.328
K2B4S1Z	0.025	0.031
K1B1S1Z	0.125	0.300
K3B5S1Z	0.075	0.154

Table B.17. Variance explained and factor eigenvalues for booklet 4.

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings ^a
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	1.984	9.918	9.918	1.327	6.637	6.637	1.052
2	1.630	8.148	18.066	0.958	4.789	11.6	0.825
3	1.429	7.146	25.212	0.908	4.538	15.964	0.828
4	1.393	6.963	32.175	0.708	3.540	19.504	0.818
5	1.294	6.470	38.645	0.605	3.027	22.532	0.773
6	1.197	5.985	44.630	0.515	2.575	25.107	0.648
7	1.093	5.466	50.096	0.421	2.107	27.214	0.564
8	1.063	5.317	55.413	0.356	1.781	28.995	0.482
9	1.007	5.033	60.446	0.325	1.624	30.619	0.693
10	0.980	4.902	65.349				
11	0.873	4.363	69.712				
12	0.827	4.133	73.845				
13	0.778	3.890	77.734				
14	0.748	3.741	81.475				
15	0.722	3.608	85.083				
16	0.681	3.405	88.488				
17	0.641	3.205	91.693				
18	0.620	3.102	94.795				
19	0.526	2.628	97.423				
20	0.515	2.577	100.000				

Table B.18. Factor matrix for booklet 4.

	Factor								
	1	2	3	4	5	6	7	8	9
K2B3S1K	0.481								
K3B6S1K	0.409								
K3B2S1O	0.391	0.317			-0.306				
K3B2S2K	0.370								
K2B7S2O	0.334								
K2B8S1O	0.314								
K1B6S1K									
K3B4S1O		0.386		0.328					
K2B6S2O		0.308	0.305						
K3B8S1O		0.419	0.645						
K2B5S2O			0.350						
K1B7S1K	0.341			-0.409					
K3B3S1O				0.349					
K1B8S2O					0.460				
K3B1S1O						0.419			
K1B1S1Z						0.323			
K2B3S1Z			-0.304				0.330		
K3B5S1Z									
K2B2S1Z								0.305	
K2B4S1Z									

Table B.19. Pattern matrix for booklet 4.

	Factor								
	1	2	3	4	5	6	7	8	9
K2B3S1K	0.606								

Table B.19 Pattern matrix for booklet 4.(cont.)

	Factor								
	1	2	3	4	5	6	7	8	9
K3B6S1K	0.406								
K3B2S2K	0.401								
K1B8S2O	0.325								
K3B4S1O		0.487							
K2B5S2O		0.458							
K2B6S2O		0.397							
K3B8S1O			0.799						
K1B7S1K				-0.614					
K1B1S1Z				-0.332					
K3B3S1O				0.332					
K3B2S1O					-0.616				
K1B6S1K					0.362				
K3B1S1O						0.558			
K2B3S1Z							0.544		
K2B2S1Z								0.408	
K2B8S1O									
K2B4S1Z									
K2B7S2O									0.614
K3B5S1Z									

Table B.20. KMO and Bartlett's test for booklet 5.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.508
Bartlett's Test of Sphericity	Approx. Chi-Square	312.696
	df	190
	Sig.	<0.001

Table B.21. Communalities for booklet 4.

	Initial	Extraction
K2B1S1K	0.143	0.256
K3B1S2K	0.206	0.569
K2B7S1K	0.251	0.631
K2B8S1K	0.124	0.552
K2B6S1K	0.084	0.107
K2B5S2O	0.193	0.502
K2B6S2O	0.106	0.504
K2B7S2O	0.222	0.320
K1B8S2O	0.134	0.269
K3B8S1O	0.224	0.430
K3B1S2O	0.099	0.154
K3B2S2O	0.127	0.152
K3B2S3O	0.181	0.491
K3B3S2O	0.123	0.220
K3B4S2O	0.079	0.112
K1B3S1Z	0.117	0.229
K1B4S1Z	0.137	0.502
K2B8S3Z	0.183	0.702
K2B5S1Z	0.140	0.370
K3B2S1Z	0.192	0.402

Table B.22. Variance explained and factor eigenvalues for booklet 5.

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.019	10.096	10.096	1.401	7.005	7.005	1.028
2	1.779	8.897	18.993	1.270	6.348	13.353	1.112
3	1.371	6.855	25.847	0.845	4.227	17.580	0.899
4	1.344	6.718	32.566	0.769	3.844	21.424	0.621
5	1.253	6.266	38.832	0.681	3.404	24.828	0.874
6	1.230	6.152	44.984	0.676	3.380	28.208	0.754
7	1.158	5.790	50.773	0.581	2.906	31.115	0.701
8	1.089	5.446	56.220	0.467	2.333	33.448	0.726
9	1.052	5.261	61.481	0.422	2.108	35.555	0.845
10	1.007	5.033	66.514	0.365	1.825	37.380	0.550
11	0.929	4.647	71.161				
12	0.838	4.188	75.349				
13	0.817	4.086	79.435				
14	0.770	3.852	83.287				
15	0.712	3.561	86.848				
16	0.609	3.047	89.894				
17	0.590	2.949	92.844				
18	0.498	2.488	95.331				
19	0.480	2.401	97.733				
20	0.453	2.267	100.000				

Table B.24 Pattern matrix for booklet 5.(cont.)

	Factor									
K3B4S2O										
K2B7S1K	0.792									
K2B7S2O	0.536									
K3B3S2O										
K2B8S3Z		0.832								
K2B5S1Z			0.595							
K3B2S1Z				0.610						
K3B8S1O				0.573						
K2B6S1K										
K1B4S1Z					0.709					
K2B6S2O					-0.706					
K2B1S1K					0.316					
K2B8S1K						0.741				
K3B2S3O							0.662			
K1B8S2O							0.338			
K3B1S2O										
K2B5S2O		0.324						0.383		
K1B3S1Z								0.383		

Table B.25. KMO and Bartlett's test for booklet 6.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.639
Bartlett's Test of Sphericity	Approx. Chi-Square	336.468
	df	190
	Sig.	<0.001

Table B.26. Communalities for booklet 6.

	Initial	Extraction
K3B4S2K	0.253	0.448
K2B7S2K	0.077	0.164

Table B.26 Communalities for booklet 6. (cont.)

	Initial	Extraction
K1B8S1K	0.108	0.378
K3B8S1K	0.131	0.161
K2B4S1K	0.072	0.069
K3B1S2O	0.196	0.277
K3B2S2O	0.239	0.459
K3B2S3O	0.112	0.189
K3B3S2O	0.124	0.206
K3B4S2O	0.093	0.264
K2B8S2O	0.158	0.467
K3B5S1O	0.130	0.550
K3B7S1O	0.152	0.257
K3B6S1O	0.074	0.125
K2B8S3O	0.231	0.484
K3B1S1Z	0.190	0.287
K3B2S2Z	0.224	0.469
K2B5S2Z	0.251	0.490
K2B6S2Z	0.049	0.036
K3B3S1Z	0.129	0.209

Table B.27. Variance explained and factor eigenvalues for booklet 6.

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.620	13.099	13.099	1.981	9.904	9.904
2	1.683	8.413	21.512	1.089	5.447	15.351
3	1.349	6.747	28.259	0.680	3.401	18.752

Table B.27 Variance explained and factor eigenvalues for booklet 6. (cont.)

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
4	1.261	6.307	34.566	0.622	3.111	21.864
5	1.167	5.836	40.402	0.499	2.494	24.358
6	1.107	5.533	45.935	0.440	2.202	26.560
7	1.080	5.402	51.337	0.374	1.872	28.432
8	1.061	5.304	56.641	0.303	1.515	29.947
9	0.997	4.986	61.627			
10	0.931	4.654	66.281			
11	0.890	4.452	70.733			
12	0.819	4.093	74.827			
13	0.773	3.867	78.694			
14	0.747	3.733	82.426			
15	0.706	3.528	85.954			
16	0.674	3.372	89.326			
17	0.595	2.973	92.300			
18	0.561	2.806	95.106			
19	0.515	2.575	97.681			
20	0.464	2.319	100.000			

Table B.28. Factor matrix for booklet 6.

	Factor							
	1	2	3	4	5	6	7	8
K2B8S3O	0.516				-0.354			
K3B2S2Z	0.477			-0.308			-0.309	
K3B1S2O	0.439							

Table B.28 Factor matrix for booklet 6. (cont.)

	Factor							
	1	2	3	4	5	6	7	8
K3B1S1Z	0.410							
K3B8S1K	0.352							
K3B3S1Z								
K2B6S2Z								
K2B5S2Z		0.608						
K3B4S2K	0.383	0.480						
K3B2S2O	0.411	-0.424						
K3B7S1O		0.344	0.301					
K2B8S2O	0.317		0.537					
K3B5S1O	0.360			0.490		0.359		
K3B3S2O								
K2B4S1K								
K1B8S1K					0.326			
K3B2S3O								
K3B6S1O								
K3B4S2O								
K2B7S2K								0.308

Table B.29. Pattern matrix for booklet 6.

Pattern Matrix	
Rotation failed to converge in 25 iterations. (Convergence = 0.000).	

Table B.30. KMO and Bartlett's test for booklet 7.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.488
Bartlett's Test of Sphericity	Approx. Chi-Square	212.464
	df	190
	Sig.	0.126

Table B.31. Communalities for booklet 7.

	Initial	Extraction
K1B8S2K	0.060	0.172
K2B6S2K	0.131	0.258
K2B2S1K	0.185	0.394
K2B5S2K	0.143	0.379
K3B3S2K	0.085	0.166
K2B8S2O	0.155	0.286
K3B5S1O	0.143	0.547
K3B7S1O	0.122	0.344
K3B6S1O	0.135	0.331
K2B8S3O	0.052	0.161
K1B2S1O	0.136	0.542
K1B1S1O	0.122	0.486
K1B3S1O	0.149	0.681
K1B4S1O	0.102	0.252
K1B5S1O	0.054	0.066
K2B7S2Z	0.119	0.473
K3B2S3Z	0.054	0.098
K1B8S1Z	0.094	0.185
K3B1S2Z	0.101	0.213
K3B4S2Z	0.198	0.494

Table B.32. Variance explained and factor eigenvalues for booklet 7.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumu- lative %	Total	% of Variance	Cumu- lative %
1	1.751	8.755	8.755	1.174	5.870	5.870
2	1.647	8.233	16.987	1.022	5.112	10.982

Table B.33 Factor matrix for booklet 7. (cont.)

	Factor									
	1	2	3	4	5	6	7	8	9	10
K3B6S1O		0.432								
K3B4S2Z	0.345	0.420								
K2B2S1K	-0.337	0.400								
K2B6S2K		0.378								
K2B8S2O	-0.321	0.357								
K1B2S1O			0.498			-0.321				
K3B5S1O	0.304		-0.470			0.302				
K2B7S2Z			0.355		0.300					
K2B5S2K					0.466					
K1B1S1O	0.384					0.426				
K3B7S1O							0.309			
K3B1S2Z							-0.301			
K1B5S1O										
K3B2S3Z										
K1B4S1O								0.319		
K3B3S2K										
K1B8S2K										
K2B8S3O									0.323	

Table B.34. Pattern matrix for booklet 6.

Pattern Matrix
Rotation failed to converge in 25 iterations. (Convergence = 0.001).

APPENDIX C: CFA MODEL ERROR

Confirmatory factor analysis for pilot test results cannot be provided since the model error. The error is shown in Appendix C.



i The model is not admissible: lavaan WARNING: some estimated ov
variances are negative

Figure C.1. CFA model error.