8-1-2018

Assessment of Multiple-Form Structure Designs of Multistage Testing Using IRT

Hanan M. AlGhamdi
University of Denver

Assessment of Multiple-Form Structure Designs of Multistage Testing using IRT

_____

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

_____

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

_____

by

Hanan M. AlGhamdi

August 2018

Advisor: Dr. Kathy E. Green

Author: Hanan M. AlGhamdi
Title: Assessment of Multiple-Form Structure Designs of Multistage Testing using IRT
Advisor: Dr. Kathy E. Green
Degree Date: August 2018

Abstract

The current study investigated multistage testing (MST) as an alternative to classical linear testing (CLT) for the General Aptitude Test (GAT). The aim was to assess the effects of two assembly methods (narrow vs. wide range—NR vs. WR), two routing methods (Defined Population Intervals—DPI— and the Approximate Maximum Information method—AMI), and two panel structures (two-stage and three-stage) on precision of ability estimates and accuracy of classification for both sections of the GAT (Verbal and Mathematics). Thus, eight conditions were examined and compared: 2 (assembly conditions) * 2 (panel structures) * 2 (routing methods).

The dataset that included a sample of 9,108 examinees was obtained from the National Center of Assessment, Saudi Arabia. The MST designs were evaluated with the criteria that the more accurate condition was the condition with the smallest standard error mean for ability estimates, and the highest agreement percentage of classification between CLT and MST.

Findings revealed trivial differences in the estimated ability and standard error means among all conditions, but the design influenced the correlations between MST and CLT ability estimates. The NR and the WR condition performed equally regarding accuracy of ability estimate and classification. The performance of the DPI and AMI were similar in precision of ability estimates, but the DPI performed better than AMI regarding classification accuracy in all conditions. The results indicated that the number

of stages was important. The correlation coefficients between the examinees' scores on MST-3Stage conditions and CLT were higher than the coefficients between examinees scores on MST-2Stage conditions and CLT.

Overall, MST can be an appropriate alternative to CLT and CAT when the MST designs are structured well using an optimal item pool. Factors such as assembling and routing methods did not have a substantial impact on the accuracy of ability estimates. That means there is flexibility to use either method—a simpler method would be as effective as a complex method. The number of stages had some impact on the precision of estimations; however, it is possible that increasing the number of items in the second stage MST-2Stage can compensate for differences. Two main recommendations from this study were: (a) the item pool should be satisfactory in MST regarding the coverage of content and range of item difficulty, and (b) the MST design with the simpler method and simpler panel structure and the complex design can perform equally. Thus, advice is to use a simpler approach and reduce effort and cost.

The main limitation of the current study was the small size of the item pool and the lack of hard and easy items. For future research, studies that compare the current combinations of various factors in different conditions of MST, using an optimal item pool, is needed to enhance the results. The influence of other factors, such as different panel structures of MST (1-2-3, 1-2-3-4), different routing, and other cut-scores can be examined to identify the optimal condition for MST and for GAT.

Keywords: Multistage Testing, Routing methods, Assembly methods, DPI, AMI, IRT-2PL, General Aptitude Test.

Acknowledgements

Final reflections on this dissertation bring me here, to express my immense gratitude to the many generous people who have helped and supported me throughout my academic journey. It would not have been possible to reach this place of completion without God's blessing, nor my parents' guidance, and words do not come close to conveying how grateful I am to my wonderful husband, for his support and inspiration.

I acknowledge with deep appreciation that this accomplishment could not have been achieved without guidance from my advisor, Dr. Kathy Green, whose belief in me was a reliable source of inspiration, and I consider it an honor to be her student. I would also like to thank the rest of the RMS faculty, providing me with clear direction, and with whom it was an absolute pleasure to work. I sincerely thank my committee members, which have enhanced my dissertation. Additionally, I would like to thank my friends in the RMS program, Dareen AlZahrani, Lilian Chimuma and Kawanna Bright for their help in the final editing of my dissertation. Special thanks to my dearest friend, Richa Ghevarghese, for her incredible words and support during our academic journey together.

I must express my sincere thankfulness to the Ministry of Education and the National Center of Assessment, their sponsorship of my education in the US was instrumental in the evolution of my academic career. Special thanks to Dr. Ali Asseri for guiding me all the way and helping me to shape my career path.

Finally, I owe my deepest gratitude to my lovely parents, sisters, and brothers, for their unconditional love, care, guidance, and support; those gifts have enriched my life more than they will ever know.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction and Literature Review

Introduction

Interest in large-scale assessment as it relates to production of an educated, available workforce has increased over the past century. The purpose of large-scale assessment is to provide extensive information about various populations. It helps stakeholders such as businesses and educators to improve systems and policies. The assessment of systems in education, healthcare, or other public services are desirable at the national and international levels. Large-scale assessment can be used to help ground inferences about large populations and conclusions about services. Examples of large-scale assessment in education include the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) that focus on comparisons across countries.

One of the important purposes of large-scale assessment in education or in other fields is to provide basic information such as average proficiency and percentages of individuals at or above certain benchmarks which can be tied to performance in varied jobs. The organizations' needs for assessment have been homogenized. That homogenization encourages the extensive use of large-scale assessment and consequently the costs associated with test development, administration of the assessment, and examinee time generally become expensive. Therefore, looking for assessment techniques that reduce both the cost and the amount of time that an individual spends on

the assessment has been a significant consideration from an engagement and data quality perspective (Yan, von Davier, & Lewis, 2014).

A technique that can lead to a revolution in assessment methods is adaptive testing as an attempt to reduce measurement error without increasing test length for individual examinees. At the same time, this technique creates new challenges for many testing organizations to create well-established procedures for producing multiple computerized test forms (Luecht & Nungester, 1998).

In the scope of tests and measurements, adaptive tests have been developed to enhance measurement precision and efficiency while reducing examinee burden. The purpose of adaptive testing is to achieve the most precise estimates of person proficiency with the shortest test and this can be obtained by matching item difficulty and person ability. For this purpose, adaptive test designs, either computerized adaptive tests (CAT) or multistage tests (MST) use adaptation algorithms; where CAT has multiple adapting points at the item level, the MST has fewer adaptation points that occur between stages and/or between items in some MST designs (Kim & Moses, 2014; Yan et al., 2014).

CAT and MST are designed to be neither too hard nor too easy for the examinee, so the test closely fits the person's ability. CAT and MST work better in terms of minimizing error than non-adaptive tests, such as classical linear tests (CLT), which are known as paper-and-pencil tests or computer-based linear tests (CBT,) which use no adaptation algorithms (Zheng & Chang, 2015). The adaptation procedures aim to reduce the error in estimating item parameters and the person's ability level while reducing the length of the test (Glas, 1988; Zheng & Chang, 2015).

Multistage testing has seen increased interest as an alternative to both CLT and CAT (Kim, Chung, Park, & Dodd, 2013). It is assumed that MST is more effective with tests that aim to measure a wide range of ability "theta" where the test is administered at the level of item sets or modules (Han, 2013). An MST is designed with items selected from an item pool calibrated before the test administration, which benefits both developers and examinees: (a) MST allows developers to maintain a higher degree of control over the content balance and quality of test structure and administration, (b) MST also allows examinees to review their item responses within each module/block compared to CAT (Yan, von Davier, & Lewis, 2014).

CAT designs present difficult challenges and controversial trade-offs. They need to be short in length with specific requirements of extensive content coverage, measurement specifications, and test security. Satisfying all these requirements while meeting the same test standards as paper test increases the difficulty and the cost for test design (Yan, von Davier, & Lewis, 2014). In the mid-2000s, a series of ongoing research efforts were initiated to establish test designs with higher flexibility in infrastructure, and that are essentially more effective. Consequently, several simulation studies and pilot tests were conducted on MST, using new content and statistical characteristics to generate tests with desirable accuracy and efficiency in measurement.

Lately, the MST framework has been implemented in several large-scale assessment applications such as achievement tests, educational admissions tests, diagnostic tests, and medical licensure tests. Yan, von Davier, and Lewis (2014) presented some examples of transition from a CLT or CAT to an MST structure. More

information about the transition can be found in chapters 21, 22, and 23 in Yan, von Davier, and Lewis (2014). One example is the development of the licensing examination of Certified Public Accountants (CPA) by the American Institute of Certified Public Accountants (AICPA). The Uniform CPA exam is the only nationally standardized exam that is required for licensure in all 55 United States' jurisdictions. The CPA aims to provide a reasonable guarantee that examinees who passed the CPA have the necessary knowledge and skills for initial licensure in the protection of the public interest (Breithaupt, Ariel, & Hare, 2014). In 1995, the AICPA Board of Examiners established a research program to evaluate the possibilities of computerization via MST; thereafter, a series of studies on MST were published in AICPA technical reports and in scholarly journals. The numerous studies on MST designs recommended the administration of MST for the computerized CPA. Example of these studies is the study conducted by Breithaupt, Ariel, and Hare (2010) that examined the effect of different MST designs and administration models with item bank development on maintaining the equality and security of the test. Also, Chuah, Drasgow, and Luecht (2006) evaluated the adequacy of item parameter estimates in terms of accuracy of ability estimates and classification using different sample sizes.

Another example of the use of MST is the transition of a K-12 assessment, the Educational Records Bureau (ERB) Comprehensive Testing Program 4 (CTP4). The bureau converted from LCT to MST. The ERB exam is not mandatory, but it aims to provide its membership with assessments, programs, and services to inform instructional and curriculum decisions. For transition purpose, the ERB used data resulting from

paper-based tests, which were online administrated, to develop an MST for CTP. This transition allows test difficulty to be better tailored to a student's ability as well as generating a simplified testing process for teachers and test coordinators (Wentzel, Mills, & Meara, 2014).

International large-scale assessment programs have also become interested in MST. Recently, real-world and simulation data have been used to conduct studies with the intention to use MST designs for the purpose of assembly and delivery of the revised version of the graduate record examinations (GRE), isolated within the verbal and quantitative sections. The GRE is globally used for graduate admission, and consists of three sections: analytical writing, verbal reasoning, and quantitative reasoning; these exams are developed by the Educational Testing Service (ETS, 2011).

Another large-scale international study is the Program for International Assessment of Adult Competencies (PI-AAC). The PI-AAC is a complex assessment that aims to assess and compare the basic skills and competencies of adults between ages 16-65; these are skills considered to be essential for success and participation is used in 27 countries (Chen, Yamamoto, & von Davier, 2014).

Statement of the Problem

In order to design an MST, many decisions must be made. Different questions are asked and need to be answered prior to the realization of an MST. Questions regarding test design and scoring include: How long must the test be? How many stages are needed? How many modules are sufficient per stage? How many paths need to be provided between modules? What routing rules should be set between modules? What is

the desired difficulty level for each routing stage and module? How will the test be scored? These types of questions must be considered for each test, stage, module, path level, routing, and scoring method.

The wide interest in the use of MST leads test developers to address the above listed numerous questions related to the MST structure, such as number of stages, number of modules per stage, module lengths, module difficulty levels, module difficulty ranges at each stage, assembly approaches, and routing methods (Yan et al, 2014). These are just a sample of the many structures critical to the development of high efficiency and accuracy of an MST.

Different studies have investigated the impact of various MST designs on the estimation of person ability and/or on classifying examinees into correct categories (see Education Testing Service, 2011; Yan, von Davier, & Lewis, 2014). Most studies have examined the performance of various panel structure designs, module difficulty ranges, and routing methods, separately using simulation data for dichotomous or polychotomous item pools (Kim et al., 2013; Kim & Moses, 2014; Kim, Moses, & Yoo, 2015; Zheng, Nozawa, Gao, & Chang, 2012; Zheng & Chang, 2015; Zwitser & Maris, 2015). Rarely are there studies found that have combined the examination of the influence of several factors, such as panel structures, routing methods, and assembly methods.

Thus, the current study addresses this gap in the literature and investigates the performance of various suggested panel structure designs, routing methods, and assembly approaches using real-world data from two content areas related to general aptitude, which are the verbal and quantitative sections of the General Aptitude Test (GAT). The

data were obtained from the National Center of Assessment (NCA) in higher education, located in Saudi Arabia. This study is potentially the first that attempts to design MST for the GAT using data from the administration of the CLT of the GAT (Arabic version). The results of this study provide practical information for MST design and can also be used with future studies on MST for the GAT to enhance the potential transition of the GAT from CLT to MST.

Study Purpose

The purpose of the current study was to examine and compare the performance of two assembly conditions (small vs. large differences), two routing methods (Defined Population Intervals (DPI) vs. Approximate Maximum Information method (AMI), and two panel structures (two-stage vs. three-stage) in terms of precision of ability estimates and accuracy of classification decisions (i.e., above average/average/below average) for both sections of the GAT (verbal and quantitative). Therefore, eight conditions were tested and compared for each section of the GAT: 2 (assembly condition) * 2 (panel structures) * 2 (routing methods) design.

Research Questions

The study aimed to answer the following research questions:

a) Does the performance of two assembly conditions (NR and WR) yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V/ GAT-M?

b) Does the performance of DPI and AMI routing methods yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V/ GAT-M?

c) Does the performance of two-stage MST and three-stage MST yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V/ GAT-M?

d) Do the examinees' ability scores on classical linear testing (CLT) relate to the examinees' scores on each condition of MST for both GAT-V and GAT-M?

Literature Review

The literature review includes three parts. The first part reviews general information about multistage testing including design, advantages of MST over CLT and CAT, implementations of MST, and examples of challenges in MST. A general overview of item response theory models are presented in the second section of the literature review. The last section provides a brief overview of the General Aptitude Test.

Multistage Testing

Multistage testing is a process for creating an adaptive test that can be administrated in a sequential process. The general idea of MST is to assess the examinees through a design that allows for adaption of the difficulty level of the test to the level of the examinee's proficiency or ability. Examinees are routed through a series of preassembled subsets of items and stages; thus, the testing procedure is systematically divided into multiple stages. The adaptation algorithms in MST are used in order to assign the examinees to an appropriate set of items, or what are called modules; therefore,

MST is a testing method that is intermediate between fully computerized adaptive testing and conventional linear testing (Kim et al., 2013; Kim & Moses, 2014; Zheng & Chang, 2015).

MST design. The MST includes several elements that introduce new concepts in testing. These elements are panels, modules, stages, and pathways. Panels are like test forms that can be constructed from the same item pool. Each panel (form) includes a quantified number of stages. The stages comprise different sets of items, which are called modules or blocks. The module is considered the smallest unit and includes a set of items that encompass a range of certain difficulty levels (e.g., easy, medium, or hard). The pathways link a module with another module. Within a specific panel, the examinees would take particular paths to connect a module in a particular stage to another module in the next, depending on their performance at the previous stage (Kim et al., 2013; Kim & Moses, 2014). Figure 1 illustrates different designs for MST: Design A has two stages and four modules with three paths, Design B has three stages and six modules with four paths, and Design C has three stages and seven modules with seven paths. Designs B and C have more adaptation points in comparison to Design A, which gives examinees a chance to recover in the third stage; Design A does not have this feature. Design C has more modules and paths than Design B, which is appropriate for a test that measures a wide range of ability.

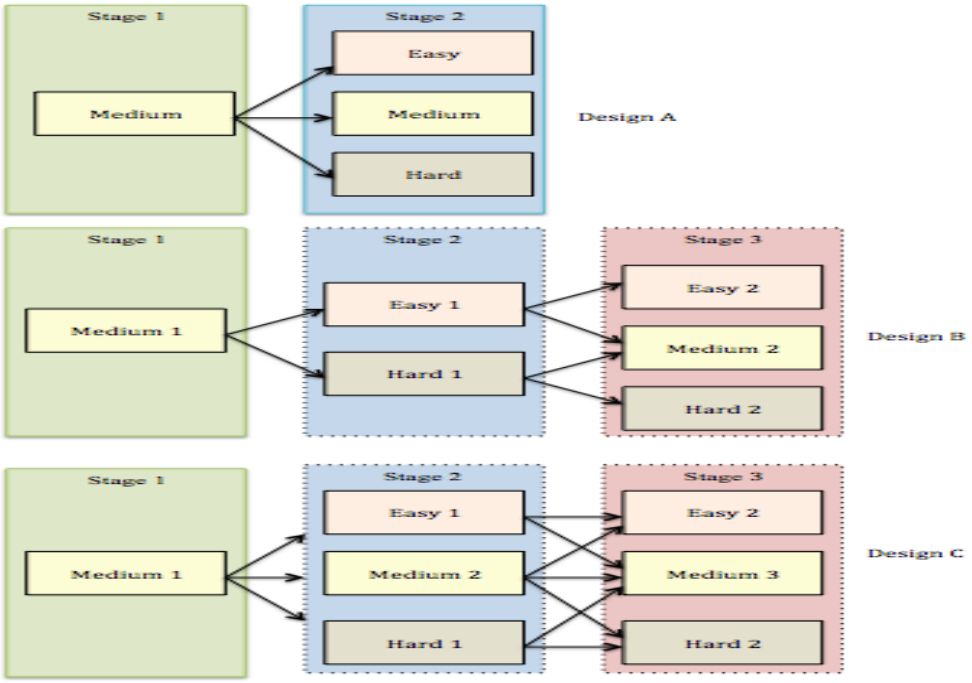

Figure 1. Three examples of different multistage testing structures

The number of panels, stages, modules, pathways, and items can vary from design to design. Different principles are relevant to the test designs, which include the purpose of the test, score usage, psychometric characteristics of items, and administration procedures (Kim & Moses, 2014). However, MST with additional stages as well as different modules in each stage permits more flexibility and a high level of adaptation for examinees. On the other hand, MST with more stages causes increased complexity in the test assembly, without necessarily increasing the measurement precision of the test. Kim and Moses (2014) pointed to a recent study by Wang, Fluegge, & Luecht (2012), which found that complex and simple MST designs performed equally well when the item bank was optimal, with high quality items that covered the range of the target ability. Therefore, the design decision should be based on the balance between a desired range of item difficulty that must be covered and the test precision that is needed. The two-stage

tests may be simple and easy to implement, but there is a high chance of routing error as a result of only one routing point. Thus, MST with three or four stages is most common in research applications (Yan, von Davier, & Lewis, 2014).

According to MST administration, all examinees are assigned to the first stage, which is called the routing stage, from one of the multiple panels/forms within the test. The routing stage usually includes one module, sometimes two, wherein the initial examinee's proficiency/ability is estimated. In the second stage, the examinees are assigned to one of the different modules with different levels of difficulty (e.g., low, middle, high modules). For instance, a MST with two stages includes three modules at the second stage. Examinees with scores that are lower than or equal to a lower cutoff score in the routing stage will be assigned to an easy module. Examinees with scores on the routing test between two cutoff scores will take a medium module. Finally, examinees with scores on the routing test higher than the upper cutoff score will continue to a hard module (Kim et al., 2013; Zwitser & Maris, 2015). Consequently, the better the performance in an earlier stage, the more difficult the follow-up stage. So, the routing stage rules the selection of the next test (Glas, 1988).

MST features. According to Lord (1980), it is assumed that multistage testing improves the measurement of the examinees' ability, but this depends on the existence of a large item pool that is calibrated. In this case, the estimation errors of item and person parameters would be reduced when the levels of item difficulty match the levels of persons' ability for different clusters of examinees. This feature makes MST an alternative to both CLT and CAT.

MST vs. CLT. Kim, Chung, Park, and Dodd (2013) reporting large-scale studies found that MST is shorter in test length than CLT, and MST provides higher or at least equally predictive and concurrent validities compared to CLT. CLT requires a larger number of items, and the examinees must respond to all items in order to achieve equal precision in the final scores; MST uses a different strategy. All examinees must take an initial set of items at the first stage. Then, based on their performances, they are routed to different modules (e.g., easy, medium, and hard) at the second stage (Yan et al., 2014). Although CLT measures average ability examinees well, it is less accurate in measuring those examinees whose ability positions are located near the higher or lower ends of the measurement scale. Accordingly, measurement precision may vary across the different levels of examinee ability (Yan et al., 2014).

Zwitser and Maris (2015) point out that the beneficial feature of adaptive tests, including CAT and MST, is a strong match between the difficulty level of items and the proficiency of examinees. This feature reduces the likelihood of undesirable response behavior in the test, such as guessing and slipping. Guessing is an unexpected correct/incorrect response that fails to show the examinee's true proficiency. In the same vein, slipping is the probability of obtaining a score of zero when at least one measured trait exists (Rupp, Templin, & Henson, 2010). Consequently, reducing the probability of guessing or slipping leads to reducing the need for estimating guessing parameters. This implies that MST can produce more parsimonious models compared to CLT.

Generally speaking, when the adaptive test designs are used with a large item pool, the model fit is expected to be better for adaptive testing than for CLT with a fixed

number of items per examinee. This implies that, despite the fact that the number of possible observations is the same in both situations, a measurement model for adaptive testing includes more parameters than the same model for CLT with a fixed number of examinees and that will help reduce the test error (Zwitser & Maris, 2015).

MST vs. CAT. Adaptive tests include two major designs: computerized adaptive tests, and multistage tests in which adaptation algorithms are used. However, the adaptive algorithms differ substantially for both designs. According to Luecht and Nungester (1998), the item selection algorithms in CAT construct every individual test form while the examinee is taking the test through the following sequential processes: iteratively administering an item, estimating a provisional score, and subsequently selecting the next item from the active item bank while considering particular statistical optimization criteria. Consequently, adaptation to the examinee's ability occurs at the item level in CAT. However, if for the first few items a capable examinee accidentally responds incorrectly, or a less capable examinee happens to guess correctly on those first few items, it is difficult to estimate their true ability levels, especially with a short test. In contrast, MST would essentially avoid this under-/over-estimation problem because the examinee's ability is estimated by the end of the first stage over multiple items (Zheng & Chang, 2015).

Similar to CAT, MST adapts the difficulty level of the test to the examinee's proficiency level. In MST, the item selection algorithms that specify adaptation to a person's ability depend on the person's cumulative performance on previous item sets. It does not depend on a single item's difficulty, as is the case with CAT. Therefore,

adaptation to the examinee's ability happens between stages of the test process. Thus, there are fewer adaptation points in MST compared to CAT (Kim & Moses, 2014).

MST allows examinees to answer the items in any order within a particular module. In contrast, using CAT, the examinees must answer each item individually in order, without skipping any, because the next item will be selected based on performance on the previous item. MST usually does not require as large a number of items as CAT requires to achieve the same standard error; however, in MST, the items in each module must meet specific criteria regarding the content, difficulty, distribution of the answer key positions (e.g. middle/extreme positions in three- or four-choice test), and length of the item (Kim & Moses, 2014; Yan, von Davier, & Lewis, 2014). In CAT, because the test is constructed during the test administration, the quality guarantee processes of the test form are limited to fulfilling particular item selection constraints. Therefore, CAT eliminates any chance to intervene in or review the criteria of both the next selected item and the individual test form for each examinee. Accordingly, the degree of trust in the test construction procedure for each test form would be insufficient in CAT because the items may be less able to cover the target range of test's content compared to MST.

The test developers or content experts cannot review each individual test form for each examinee who is taking a CAT. Thus, CAT is inefficient in combining the content specifications of the test, even with the most refined content-balancing algorithm (Kim et al., 2013; Luecht & Nungester, 1998). In CAT, the most common criticism is the test is built on-the-fly from an item pool during the test administration; therefore, content specialists are unable to review every test form for every examinee before the test is

given to verify that all test forms have satisfied all content requirements and there are not context effects, such as one item influencing the response to another item (Keng, 2008; Kim et al., 2013). Thus, it is generally believed that CAT is incapable of integrating all the content specifications of the test. In contrast, the modules in MST are arranged into stages with pre-determined routing rules; thus, the MST has a better-quality control procedure that allows content specialists to review the test content in order to ensure the content balancing for each form before test administration (Keng, 2008). Since the MST consists of a small number of distinct modules, the test developers can ensure that each module satisfies certain requirements regarding item content, item difficulty, total word count, and distribution of answer key positions. Additionally, they can prevent or avert dependencies between items in the same module (Kim & Moses, 2014). The feature of constructing the MST forms prior to test administration gives MST substantial advantages compared to CAT. Table 1 sumarizes the main differences between CLT, CAT, and MST.

Table 1.

Main Differences between CLT, CAT, and MST

Domain	CLT	CAT	MST
Adaptive points	N/A	Multiple adaption points at the item level	Fewer adaptation points between stages
Content Specification	Sufficient	Inefficient	Sufficient
Content review before test administration	Avaliable	N/A	Avaliable
Content coverage	Adequate	Inadequate	Adequate
Item response order	Answering items in any order	Answering item individually in order, without skipping any	Answering items in any order within a particular module
Chance of guessing	High	High	Low
Recovering	N/A	No chance to recover	Recover from misrouting with 3 stages and more
Test length	Long	Long	Shorter than CAT and CLT

MST implementations. Generally, there are three implementations of MST: (1) paper-and-pencil test with separate administrations, (2) computer-based test with separately timetabled administrations, and (3) continuously administered test at a single administration. The first and the second tests are less common and are statistically analyzed in a similar way. The third implementation type is the most commonly used, as well as the most challenging for analyses (Yan, von Davier, & Lewis, 2014). MST is usually used in the context of achievement testing and classification, for example, MST is used for the Uniform Certified Public Accountant and the United States Medical Licensing Examination. Computer software is usually used to conduct and assemble the

MST's panels and modules, with software such as the automated test assembly (ATA) program. The ATA program creates fixed-length parallel test forms either with or without statistical constraints. The ATA also applies optimization algorithms (e.g., linear programing, heuristics) to concurrently generate multiple panels, stages, and modules in the MST (Kim et al., 2013; Luecht & Nungester, 1998).

Briefly, the ATA optimization can be solved by using heuristic methods, such as the normalization weighted absolute deviation (NWAD), which is capable of managing complex content structures and requirements. With MST, the NWAD heuristic methods integrate specific mechanisms that can sequentially manage the assignment of modules to stages within a panel, as well as incorporate procedures for dealing with different simultaneous objective functions, such as using different target test information function (TIF) for modules or combining modules and building multiple test forms. Therefore, the total set of items for all stages in the MST accurately satisfies every content specification for a full-length test form (Luecht & Nungester, 1998).

Either the ATA program or the manual test construction can be used in the MST, but most MST studies have constructed and applied the Two-stage or Three-stage MST, which includes a single module at the first routing stage and three modules at the second and/or third stages.

MST challenges. Computerized testing encounters challenges regarding design, administration, production, and cost. Testing agencies that change the test forms from the CLT (paper-and-pencil) forms to the computerized forms recognize that computerized testing, such as CAT and MST, requires a higher level of proficiency compared to CLT

in term of design, assembling, routing, scoring, and linking methods. Along with the advantages of computerized tests, the literature on MST shows that it introduces a series of new challenges to assure the quality of all procedures. Different potential practical issues and considerations for operational MST processes can occur, from design to application. The most practical issues and psychometric considerations for operating MST are related to designing, assembling, routing methods, scaling, calibrating, linking, equating, and securing the test (Yan, von Davier, & Lewis, 2014).

Design challenges. Design issues and considerations of the MST have been discussed by large-scale studies that included a range of possible structures for MST (1-3, 1-2-2, 1-3-3, 1-2-3, 1-3-2, 1-1-2-3, 1-1-2-3-3-4, etc.), meaning the MST can be a two MST design such as 1-3 design that includes one module in the first stage and three modules in the second stage. MST design also can be a three MST design such as 1-2-2, 1-3-3, 1-2-3, 1-3-2 design that have usually one routing stage and two additional stages with different numbers of modules in the next stages. Zenisky and Hambleton (2004) concluded that the crucially important consideration in designing the MST is to determine the number and the characteristics of stages in the test as well as the number and features of modules per stage. However, the optimal measurement results for any design rely on the distribution of proficiency in the target population and the operational item bank. An example of the importance of the examinee population distribution would be that with a wide population proficiency distribution (e.g., achievement testing) more modules per stage are needed; this creates a better chance of spreading examinees out on the basis of proficiency. Another example that illustrates the significance of the item bank can be

seen when the MST design requires only two modules per stage (e.g., easy, hard) to make a pass-fail decision. In this case the item bank needs to include sufficient items within the hard and easy levels of difficulty (Yan, von Davier, & Lewis, 2014). In summary, the selection of the test design is more complex in MST than in CAT and CLT, and the test development decision must be determined by the test purpose and measurement requirements of the testing.

Assembly challenges. Test assembly methods present another challenge in MST due to the large number of possible paths. It is important to use statistical and non-statistical requirements in order to have a balanced representation of items from each content area. The assembly approach in MST is different and more complex than in CLT. In general, there are two distinctive approaches for assembly in the MST: (1) automated test assembly ATA, which was noted earlier in this study, and (2) assembly-on-the-fly, which is a new method that uses the same item selection algorithms as those used in CAT, to construct individual modules for each examinee dynamically (Yan, von Davier, & Lewis, 2014; Zheng & Chang, 2015).

There are various ATA algorithm methods that are used in MST and these methods can be applied to assemble various modules from item banks, in addition to the capability of being used to assemble panels from modules. Zheng, Nozawa, Gao, and Chang (2012) discussed three common ATA methods: (a) early test assembly methods, (b) the 0-1 programming methods, and (c) heuristic methods. The early test assembly methods relying on statistics that stem from classical test theory CTT (e.g., item difficulty and discrimination statistics) or item response theory IRT (e.g., test information

function). The 0-1 programming methods are used to optimize an objective function over a binary space, but it is challenging when several parallel forms are required with an insufficient item bank because all forms must meet all test assembly restrictions. In heuristic methods, items are selected sequentially; thus, these methods are in a way greedy since tests or forms that are assembled earlier have access to more adequate items than those forms that are assembled later. The heuristic methods do not guarantee that all constraints will be satisfied, as the 0-1 programming approach does. However, the heuristic methods are faster and do not rely on solver software like the 0-1 programming methods (Zheng et al. 2012; Zheng & Chang, 2015). However, satisfaction of both features (easy computation and guarantee all constraints) has not been simultaneously obtained. Easy computation is obtained, as in the case with heuristic methods, and the guarantee that all constraints will be met is obtained in the case of the 0-1 programming methods. The limitations of these methods introduce opportunities for research on alternative methods, such as On-the-fly Multistage Testing (OMST), which allows achievement of both features simultaneously.

Assembling parallel forms is more challenging in MST than in CLT, especially when the item bank is limited, and multiple constraints need to be fulfilled. When a test is preassembled around a few difficulty anchors, the classical MST may not provide sufficient information nor suitable ability estimates for examinees at the two extreme ends of the scale. Additionally, test security may be threatened because items are bundled together in modules and modules are bundled together in panels (Yan, von Davier, & Lewis, 2014; Zheng & Chang, 2015). Therefore, if the examinees whose abilities are

similar shared test items, or the items were disclosed on the Internet, the test overlap rate might be high among them because they are likely to receive the same panel and pathway (Yan, von Davier, & Lewis, 2014). This can be solved in OMST by borrowing a feature of CAT. In the OMST, examinees' abilities are estimated in the first stage, which is at a moderate difficulty level, and then the individual module is assembled for each examinee based on his/her ability estimate, in addition to other constraints (e.g., content coverage and item exposure rate). The OMST uses CAT methods wherein each item is administrated as long as it is selected, effectively updating the ability estimate after each item within the same module in a specific stage; this process continues until the exam is terminated. Finally, the total score of each examinee is estimated based on her/ his performances on all the administrated items. Figure 2 is an example of a design using on-the-fly methods to assemble stages in OMST as was illustrated by Zheng and Chang (2015). The great advantage of OMST over classical MST is providing sufficient information and estimating examinees at the extreme ends of the ability scale (Yan, von Davier, & Lewis, 2014).

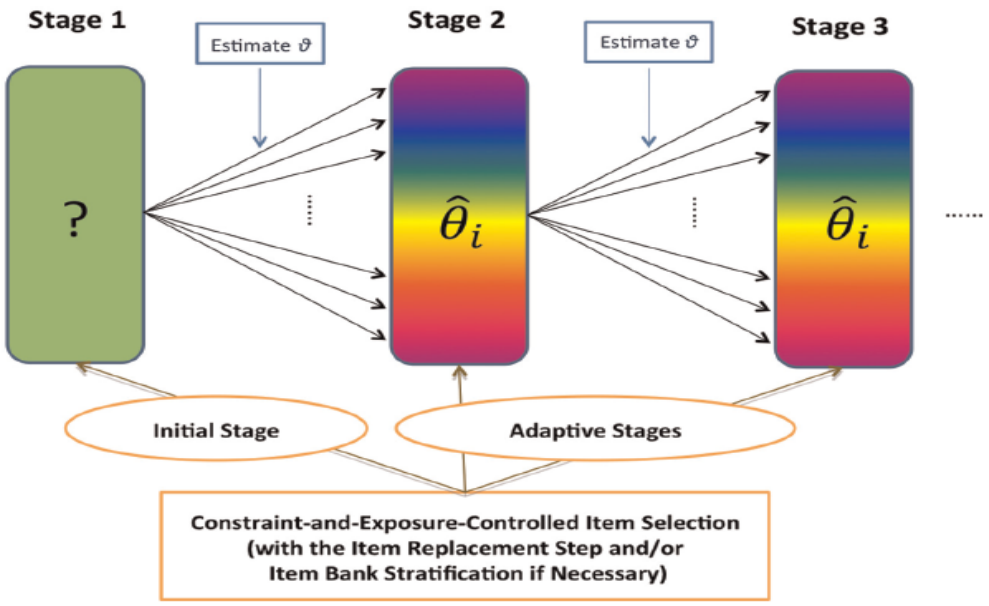

Figure 2. An example design using On-the-fly Assembling Method for MST. Adapted from "On-the fly assembled multistage adaptive testing," by Y. Zheng and H. H.Chang, 2015, Applied Psychological Measurement, 39(2), p. 106.

Routing and scoring challenges. Routing is basically a classification issue in MST; it is a process that uses special rules to route or classify examinees to different modules at each next stage, based on their score on the previous modules. The purpose of such a routing procedure is to gather information about the examinees in order to assign a module to them in the next stage. This is done by dividing the population in a certain number of subpopulations whose abilities are similar. The routing process is also important because it contributes to the results of the whole test (Yan, von Davier, & Lewis, 2014).

The two routing rules are: (1) statistical rules based on the number of correct responses and (2) dynamic rules based on the examinee's performance on items in the first module, with consideration for the properties of items in the next modules. Statistical

routing rules are easier to implement than dynamic rules, yet dynamic rules have certain advantages over statistical rules (Weissman, 2007).

The two main methods of routing examinees are norm- or criterion-referenced routing. In norm-referenced routing, the examinees are assigned to the next module based on rank order, such as assigning one third of examinees to each module, or maybe assigning 30%, 40%, 30% to easy, medium, hard modules. This proportional method of routing has been known as Defined Population Intervals (DPI) in the literature. The *DPI* approach routes a proportion of examinees to various modules at the next stage. This technique sorts examinees, who have taken the target stage test (i.e., routing or second stage), in rank order consistent with their estimated abilities. Next, it assigns a certain percentage of the examinees to each module of the subsequent stage (Kim et al., 2013). Furthermore, the examinees can be rank-ordered within each module, according to their estimated abilities and can then be routed to a subsequent-stage module. Therefore, two methods of DPI can be used. The *stage-level DPI* (SL-DPI) method is used to route examinees to the next stages' modules and the *module-level DPI* (ML-DPI) method to route examinees within each module. Another related strategy was explored by Armstrong (2004) where the modules are targeted to percentiles of the examinee population, so the strategy aims to assign examinees to modules based on percentile group membership (Yan, von Davier, & Lewis, 2014).

In a criterion-referenced routing strategy, the examinees are assigned according to their proficiency indicators that are compared to a performance-based rule. This approach matches observed performance and module assignment; therefore, it is called

the Maximizing Information (MI) approach. This approach has two options: Information-based routing vs. number-correct routing. Information-based routing assigns examinees to modules according to ability estimates that are computed by using item response theory (IRT) methods. The MI method is also called the *Approximate Maximum Information* (AMI) method to route examinees to the next stage. The AMI method uses the test information functions (TIFs) by defining the connection between the cumulative TIFs of adjacent modules (Kim et al., 2013).

A number-correct (NC) routing method reviews and sums item-level performance for an examiner, and then the scores are used for module assignment. Armstrong, Jones, Koppel, and Pashley (2004), as well as Davey and Lee (2011), compared these two criterion-referenced routing approaches and found they are quite similar in terms of results. However, the information-based approach provides a better match between examinees and modules, whereas the NC approach is simpler to explain to examinees, which is a major consideration (Yan, von Davier, & Lewis, 2014).

Even though MST has many advantages, there are also potential disadvantages of misrouting that could potentially occur due to fewer adaptation points spatially in two-stage MST. Thus, the challenge of routing is crucial because the opportunity for adaptation is limited, compared to the CAT. The impact of misrouting may include a bias in the estimation of the examinee's ability if the examinees are assigned to modules that are not matched with their ability levels. The effects of misrouting may be exacerbated when there are few items in the routing stage and the chance of guessing is high. Additionally, the misrouting impact can be substantial with modules that cover a very

narrow range of ability or with modules that have more overlap in item difficulty (Kim & Moses, 2014). However, scholars have investigated the impact of routing methods. For example, Kim et al. (2013) compared three routing methods for four panel structures. They found similarity in terms of precision and in classification decisions with the same length of test. Kim and Moses (2014) also compared the differences in examinees' scores with different conditions that included a small-difference condition with overlap in difficulty and a large-difference condition with distinction in difficulty; the results of their studies indicated minimal impacts of misrouting.

All this considered, MST has the potential to use different routing and scoring methods in combination. For example, the routing rule can be based on NC scores, while IRT estimation methods (such as maximum likelihood (ML), maximum *a posteriori* (MAP), or expected *a posterior*i (EAP) method), can be used to estimate the final score (Weissman, 2007). For MST, Lord (1980) states that it is not recommended to use NC scoring to estimate the final level of the examinee's ability because the items each examinee receives are not statistically equivalent (Kim et al., 2013). Recently, other nonparametric approaches for MST have been introduced, such as a regression tree-based approach for routing and scoring that can be used when the IRT assumptions are not met or the sample sizes are too small (Yan, von Davier, & Lewis, 2014).

Linking and equating challenges. In practical implementations of MST, it is important to ensure the comparability of MST results over time; thus, essential principles for linking and equating MSTs should be taken into consideration. The calibration and linking of tests are involved in the initial phase of data collection. First, a conventional

test is administrated to develop initial modules and routing rules in order to use them in MST administrations. Afterward, data are collected from MST administrations and the scoring rules are established, modules are built, and the cut-scores for routing are equated. These procedures are necessary to ensure the comparability of tests over time; therefore, routing rules should be comparable for different administrations (Yan, von Davier, & Lewis, 2014).

Estimation and linking are substantial requirements for practical implementation of MST. Generally, IRT is used to provide the basic framework for linking, scoring, and equating. Different approaches to estimation of parameters can be used, including: (1) concurrent calibration, (2) sequential linking, and (3) simultaneous linking. In concurrent calibration, all parameters are estimated at once for multiple administrations. Sequential linking involves separate calibrations so when the parameter estimates are found for an administration, they are never changed based on later data obtained. Simultaneous linking involves additional separate calibrations where the parameter estimates are computed simultaneously for all administrations; as a result, the modification of older parameters can occur. With a small number of administrations and items in the item bank, concurrent calibrations are usually applied with a marginal maximum likelihood (MML) estimation once. In a separate calibration, such as sequential linking and simultaneous linking, MML is applied separately to each administration. However, if the number of administrations increases, simultaneous linking is more effective than sequential linking regarding the growth of random linking errors (Yan, von Davier, & Lewis, 2014).

Nevertheless, other factors, such as item exposure, examinees' motivation, and item representation can affect the parameter estimates from administration to administration; therefore, recalibration is needed in MST. Besides MML, conditional maximum likelihood (CML) and Bayesian approaches can be applied. The advantage of CML is that no assumptions are required regarding ability distribution in the population; in addition, its use does not require a random sample (Zwitser & Maris, 2015). Glas (1988) compared two estimation methods in MST, CML, and MML methods. He found that a CML estimation method is more powerful than a MML method. However, in MST the CML method breaks down because the estimation of all item parameters simultaneously fails to work. Moreover, it is impossible to estimate parameters in the different subgroups and to calibrate the estimates on scale via the common items. Zwitser and Maris (2015) used the Rasch model in their derivations and examples to prove that CML can be applied with MST, and that the Rasch model is less restrictive in adaptive testing compared to linear testing.

Security challenges. The threat of test security is higher in MST compared to CAT because the items are set together in modules and then the modules are set together in panels. Therefore, if the items are shared among groups or disclosed on the Internet, there will be a high chance that examinees who take the same pathway as the discloser will be able to answer items correctly, regardless of their ability levels (Zheng & Chang, 2015).

The literature provides possible tools for monitoring quality and assessing test security. Various statistical indices used in CAT can also be used in MST. The most

common index is the average test-overlap rate, or the expected proportion of common items among examinees. Item exposure rate can be also calculated as the ratio between the number of items administrated and the total number of examinees. If the test-overlap rate is high, the chance of sharing information divulged by those who took the test earlier will increase. Standard deviation (SD) of test-overlap rate is another index that is important to be conducted in MST. A large SD reveals that certain groups of examinees share more common items than other groups. The SD adds more information in the MST than the mean because at times during which the mean overlap rate is the same in MST and CAT, the SD of test overlap tends to be larger in MST, which poses a risk for test security (Wang et al., 2014). Other indices, such as an item polling index and organized item theft index, which are used in CAT, can be also applied in MST.

In MST, certain quality control and test security methods are considerations that can be applied during test assembly, before test administration, and some of them are statistical methods for detecting irregularities in post-administration analysis. For example, before test administrations, some considerations can involve large item pools, multiple forms of modules, and multiple parallel panels. Additionally, the on-fly-assembly method of MST is less likely to expose items because the items are selected on the fly. Lee, Lewis, and Davier (as cited in Yan, von Davier, & Lewis, 2014) describe three statistical component procedures for monitoring quality control for MST. The three components are: (1) exploratory analyses for understanding the assessment and the examinees better to create a baseline, (2) short-term detection tools, and (3) long-term monitoring tools to assess the quality and security of the MST.

The first component involves the use of variables obtainable from the data, including response data and timing data; thus, the data must be examined at item level, module level, and test level. Also, subgroup analyses using demographic information should be conducted. This step is to ensure that items function as expected and the examinees represent the target population composition. The second component (short-term detection tools) is to detect test inconsistency in the performance on the test for each path of the MST based on residual analysis, profile analysis, or person fit analysis, which will potentially identify unexpected behaviors from regular behaviors. This analysis can be conducted across sections in the test or across items in a timed section. The pattern of module scores may be similar for different examinees. The unusual patterns may be due to cheating. The third component (long-term monitoring) is a crucial strategy for the testing industry wherein many tests are administrated per year globally.

It is necessary to investigate item and module performance over time, as well as test score distributions with summary statistics. Quality control charts, such as Shewhart and cumulative sum (CUSUM), can provide a visual examination of data with control limits to monitor the item and module performance over time in terms of accuracy and speed components. Thus, if the item or module becomes easier or less time-consuming over time, this item or module may be compromised. Moreover, investigating the differences in test scores of examinees or groups can be used to monitor examinee or group performance to demonstrate if there is an extreme change in the score. For a subgroup, the application of weighted liner mixed models to examine the data related to demographic information variables can be applied to detect the patterns in the data,

predict subgroup performance over time, and then develop prediction intervals for observations in future administration (Lee & von Davier, 2013).

Item Response Theory

Item response theory (IRT) is model-based measurement that has rapidly became a basis for assessment in the field of education, health, and social sciences. Most computerized adaptive tesing and multistage testsing are IRT-based applications. The IRT measurement models are applicable to CAT and MST for many design components that are necessary for CAT and MST adminstrations, such as routing, scoring, classification, and equating. In CAT and MST, multiple forms must be scored so each form yields comparable scores. Therefore, the item parameters must be estimated and scaled accurately and IRT accommodates the differences and provides comparable latent trait theta estimates (Weissman, 2014). IRT provides advantages over classical test theory (CTT). IRT is more psychometrically robust in creating measures of latent traits than CTT in addition to its benefit of reducing test length. The key feature of IRT over CTT that is IRT provides comprehensive descriptions of performance for each item in the test. IRT also produces item and scale indices for accuracy that freely vary across the full range of all possible scores. The bias of items and tests can be assessed in IRT by demographic subgroups while also measuring each examinee's response pattern quality and consistency (Harvey & Hammer, 1999). Therefore, IRT is widely applied to calibrate and assess items, instruments, and also to score examinees on their abilities or other latent traits. IRT methodology provides critical improvements in measurement accuracy and reliability; thus, IRT-based techniques have been used to develop educational tests, such

as the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE) (An & Yung, 2014).

Item response theory uses a logistic regression formula to estimate responses using latent characteristics of individuals and items as predictors; thus, the examinees' performances can be explained by latent traits (De Ayala, 2009; Zhang, 2013). The probability models are used in IRT to characterize the interaction between test takers and test items. The trait level in IRT is estimated based on both persons' responses and on the properties of the administered items (Embretson & Reise, 2000). For example, in the two-parameter IRT, the probability of a correct response is a function of two parameters: The distance between the person location (person ability) and the item location (item difficulty), and the differentiation among examinees' locations at different points on the continuum (item discrimination). The three-parameter IRT model (3PL) adds a third parameter: guessing, which is useful if the test is presented in a multiple-choice item format. In this case, examinees with low proficiency may select the correct response by guessing. Therefore, the 3PL addresses issues that occur when the item response function has a lower asymptote that is a nonzero value (De Ayala, 2009).

Lord in the 1950s and Rasch in the 1960s introduced IRT as an alternative model-based measurement theory (Embretson & Reise, 2000; Kean & Reilly, 2014). Additionally, IRT was developed to solve different practical testing issues, such as equating different test forms, score interpretation, and test score comparison within and between individuals (Embretson & Reise, 2000). Initially, IRT models were used for dichotomous items formats wherein there were only two possible scored values. Such

formats that fit this category are true/false or multiple choice with one correct response; this can be either the Rasch model or the two-/ three-parameter logistic model. However, later IRT models were extended to apply to other item formats, such as rating scales or the polytomous format, which was introduced by Andrich in 1978. In this item format, there are more than two possible scored values, as in Likert scales. Other advanced IRT models were also developed; for example, partial credit scoring was created by Masters in 1982 and multiple category scoring was originated by Thissen and Steinberg in 1984. These types of IRT models allow multiple responses with unequal scored values (Embretson & Reise, 2000; Harvey & Hammer, 1999). Additionally, IRT methodology can be applied to unidimensional models as well as multidimensional IRT models.

IRT models. Persons and items are located on the same continuum in IRT; therefore, IRT models differ based on the number of parameters being used to model the responses to each item and all examinees. It is assumed that the relationship between the trait θ and the observed item response varies across the range of θ scores as a function of the item parameters (Harvey & Hammer, 1999). Consequently, there are different IRT models related to different item parameters. The most common IRT models used in psychological and educational measurement come down to three models: The one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model. Graded response, Guttman, and Mokken models are also other models that are frequently used with rating scales. The graded response model deals with ordered polytomous categories such as letter grading, A, B, C, D, and F, or attitude rating scales that range from strongly disagree to strongly agree (Samejima, 1997). Mokken and

Guttman scaling are used to evaluate whether the items measure the same underlying concept (unidimensional scale) based on the asssumption that the items are hierarchically ordered in their difficulty so it is assumed that respondents who answer a difficult item are more likley to answer an easy item. The key difference between the two models is the probabilistic nature in a Mokken model ranging between 0 and 1.00 (DeJong & Molenaar, 1987).

One-parameter model (Rasch model). The 1PL, or the Rasch model, is the simplest of IRT models. It requires only a single item parameter to model the response process, which is the difference between the person parameter (b) and the item difficulty (d) parameter (Harvey & Hammer, 1999). The difficulty parameter is the location of items on the latent continuum that represents the construct. The upper end of the continuum indicates greater proficiency on the target trait than the case on the lower end. Therefore, by having both person and items located on the same continuum common scale, it is conceivable to predict the person's response on an item (De Ayala, 2009). It is possible to predict the probability of a response as a function of both person and item locations; however, the relationship between the differences and item responses depends on the nature of the dependent variable, which is modeled as either log (odds) or probability.

In odds ratio format, the ratio of the probability of success for person *s* on item *i* (P_{is}) to the probability of failure (1- P_{is}) would be as follows (Embretson & Reise, 2000):

$$\text{In}\,[P_{is}/(1-P_{is}] = \theta - \mathrm{d_i} \quad (1)$$

The Rasch model is one that is simple, and that can be applied to any item if the item difficulty is known. In this model it is assumed that when the trait level is equal to item difficulty, the odds of success will be zero that yield an odds of 1.0 (0.50/0.50). This means that a particular person has an equal chance to succeed or fail on any given item. Consequently, the chance of success on a particular item increases as the trait level exceeds item difficulty (Embretson & Reise, 2000). In another form of the Rasch model, or what is called the one-parameter logistic model (1PL), the dependent variable is predicted as a probability rather than as log odds. The probability that person *s* passed item *i* is predicted from the combination of item difficulty and person ability, which has a nonlinear relationship with the dependent variable; therefore, the logistic model would be modeled as follows:

$$P(X_i = 1|\theta, d_i) = \frac{e^{(\theta - d_i)}}{1 + e^{(\theta - d_i)}} \qquad (2)$$

Where $P(X_i = 1|\theta, d_i)$ is the probability of the response of 1, θ is the person location (person ability), d_i is the item's location (item difficulty), and *e* is a constant value equal to 2.7183…, according to Equation (1), the probability of a response of 1 on item *i* is a function of the distance between person ability (θ) and item difficulty (d). The dependent variable as a probability is more familiar than the log odds ratio expression (Embretson & Reise, 2000).

By implication, in the 1PL, all items in the test show an item characteristic curve (ICC) of identical shape (S-shaped curve). For the Rasch model, the predicted Rasch ICC is not as steep as the observed response function/ the empirical trace line. Therefore, to match the empirical trace line, the slope of the ICC needs to increase. This can be

completed by multiplying the exponent$(\theta - d_i)$ by a, which is the slope (item discrimination (a)); therefore, Equation (2) becomes:

$$P(X_i = 1|\theta, d_i) = \frac{e^{a(\theta - d_i)}}{1+e^{a(\theta - d_i)}} \qquad (3)$$

By adding (a) which is related to the slope of the logistic regression line, any variation in α leads to a change in the line's slope. However, the difference between the IPL and Rasch models lie in the constant value for this α, which is item discrimination. In the Rasch model, this value is constant at 1.0, whereas it does not have to be 1.0 in 1PL. In short, the Rasch model assumes that all items have the same discrimination ability, which is equal to 1.0; the IPL model however, assumes the discrimination parameter for all items is constant, and can be any value, whether 1.0 or other values. Mathematically, the two models are equivalent. Philosophically, the 1PL model puts more emphasis on fitting the data given the model's constraints as accurately as possible. In the Rasch model, the focus is to construct the latent variable of interest (De Ayala, 2009).

Two-parameter model. The one-parameter IRT model provides important information about performance using only one item property, which is item difficulty; it ignores the use of other item features, such as item discrimination. In the 1PL, it is assumed that all test items have identically shaped ICCs, in which case they would have the same discrimination. In empirical practice, the ICCs differ for each item due to some items having stronger or weaker relationships with the underlying latent construct. Therefore, the two-parameter logistic model (2PL) allows the ICCs to have different slopes, which means different items have different discrimination parameters (Harvey & Hammer, 1999). Items with higher "a" parameters are strongly related to the latent trait θ

and so provide more information regarding theta. This leads to a decrease in the standard error for estimated person ability, which increases measurement accuracy.

The 2PL model places emphasis on incorporation of how well an item discriminates among examinees located at different points on the latent trait continuum. Thus, the probability of a response passing a particular item is a function of not only the distance between person location and item location, but also on item differentiation among examinees located at different points on the latent continuum. The model that takes item discrimination into account actually relaxes the constraint that items should have the same common slope. Consequently, the ability of the 2PL model to achieve model fit in several situations is higher than the1PL or Rasch models. This is due to the latter model's allowing the slope to vary across items (De Ayala, 2009). In doing so, the 1PL equation (2) becomes the 2PL model by multiplying the exponent $(\theta - \mathrm{d_i})$ by "*a*", as follows:

$$\mathrm{P}(\mathrm{X_i} = 1|\theta, \mathrm{d_i}, a_i) = \frac{e^{a_i(\theta - \mathrm{d_i})}}{1 + e^{a_i(\theta - \mathrm{d_i})}} \quad (4)$$

The subscript on *a* indicates that each item *i* has its own discrimination parameter. The 2PL equation (4) may include a scaling constant *D* to become:

$$\mathrm{P}(\mathrm{X_i} = 1|\theta, \mathrm{d_i}, a_i) = \frac{e^{Da_i(\theta - \mathrm{d_i})}}{1 + e^{Da_i(\theta - \mathrm{d_i})}} \quad (5)$$

Item discrimination values vary from $-\infty \; to \; \infty$; however, the reasonable values range from 0.8 to 2.5, and an item with negative *a* should be eliminated because its performance is inconsistent with the model (De Ayala, 2009).

Three-parameter logistic model. In 1PL and 2PL models, it is assumed the lower asymptote of the ICCs is fixed at zero, meaning the expected proportion of correct

response that would be expected from examinees who have very low ability. However, this factor may vary across items, especially in multiple-choice item formats. Individuals with low-proficiency and who do not know the correct response may potentially guess the correct response, which produces nonzero rates of correct responses to difficult items. This is a significant factor that needs to be addressed in the measurement models. The three-parameter model represents a third parameter, which is guessing, or the pseudo-chance parameter, "c", which allows the lower asymptote of the ICCs to adopt nonzero values for minimum values. Therefore, the three-parameter logistic model equation would be as follows, after extending the 2PL model equation (5) to include the third parameter "c":

$$\mathrm{P}(\mathrm{X_i} = 1|\theta, \mathrm{d_i}, a_i, c_i) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - \mathrm{d_i})}}{1 + e^{Da_i(\theta - \mathrm{d_i})}} \quad (6)$$

where $\mathrm{P}(\mathrm{X_i} = 1|\theta, \mathrm{d_i})$ is the probability of the response of 1, θ is the person location, $\mathrm{d_i}$is the item's item difficulty parameter, a_i is the item's discrimination parameter (slope of ICC), c_i is the pseudo-chance parameter (guessing effect), *D* is a scaling factor, and *e* is a constant value equal to 2.7183. In items that have values of c_i greater than zero, the probability of correct responses at d_i become greater than 0.50. Thus, the relationship between the item discrimination and the slope depends on the item guessing parameter.

IRT Assumptions. IRT models involve stronger assumptions than other measurement models. Violation of these assumptions may yield difficulties in the results' interpretation because the latent ability estimates produced by calibrating the data to the IRT model assume that the data fit the model well (Kean & Reilly, 2014). Two main assumptions are required for item response theory: (a) the item characteristic curves

(ICCs) or item response functions (IRF) have S-shaped curves, and (b) local item independence (Embretson & Reise, 2000; Harvey & Hammer, 1999).

The first assumption pertains to the ICCs' specific form. The ICC's shape illustrates the relationship between the changes in trait level and the changes in the probability of a particular response, which are the most important relationships in IRT existing between the underlying latent construct and the response to each item in the test. The ICC is considered a two-dimensional scatterplot of theta (X-axis), by item response probability (Y-axis) (Harvey & Hammer, 1999). The ICCs for dichotomous items (e.g., correct/ wrong) regress the probability of item success on trait level. In contrast, the ICCs for polytomous items (e.g., Likert scale) regress the probability of item responses in each category on trait level (Embretson & Reise, 2000). The forms of ICCs must be S-shaped, which plots the probability of a correct response related to person ability and item parameters. Therefore, the shape of the ICC is not only impacted by trait level but also by item characteristics, such as item location, item discrimination, and item lower asymptote. Item discrimination is related to slope and shows how rapidly the probability changes with trait level. Guessing the correct response influences the item lower asymptote; thus, the lower asymptote will be greater than zero. Embretson and Reise (2000) illustrated the impact of item parameters on three items through two figures. Figure 3 shows three items with different locations on the trait scale where item 1 is easier than items 2 and 3 because it requires lower trait ability than the others. Finally, Figure 4 shows three different items with different lower asymptotes, with the probability of correct responses to item 3 that never drop to zero.

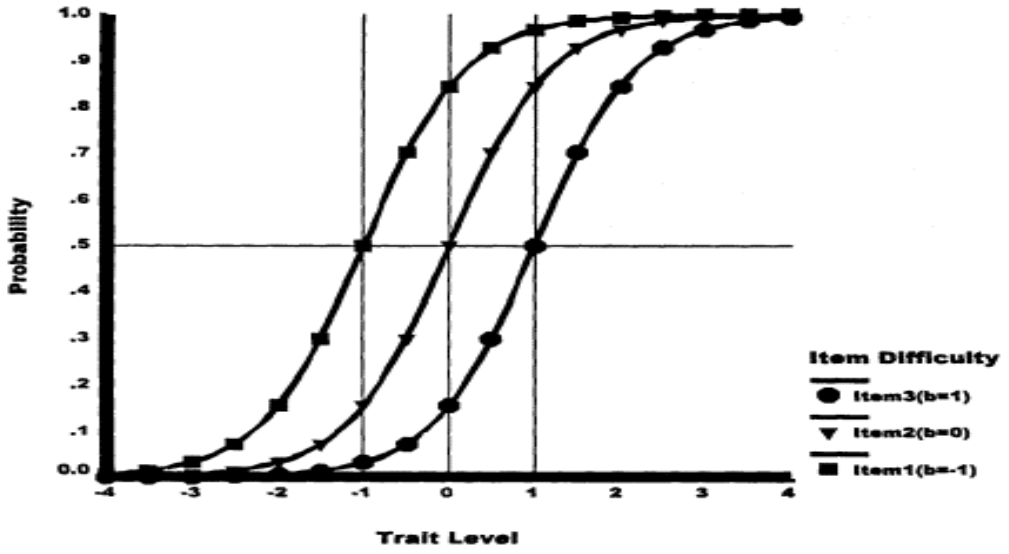

Figure 3. The ICCs for three items with different difficulty levels. Adapted from Item response theory for psychologists (p. 46), by Embretson, S., & Reise, S., 2000, Mahwah, NJ: L. Erlbaum Associates. Copyright 2000 by L. Erlbaum Associates.

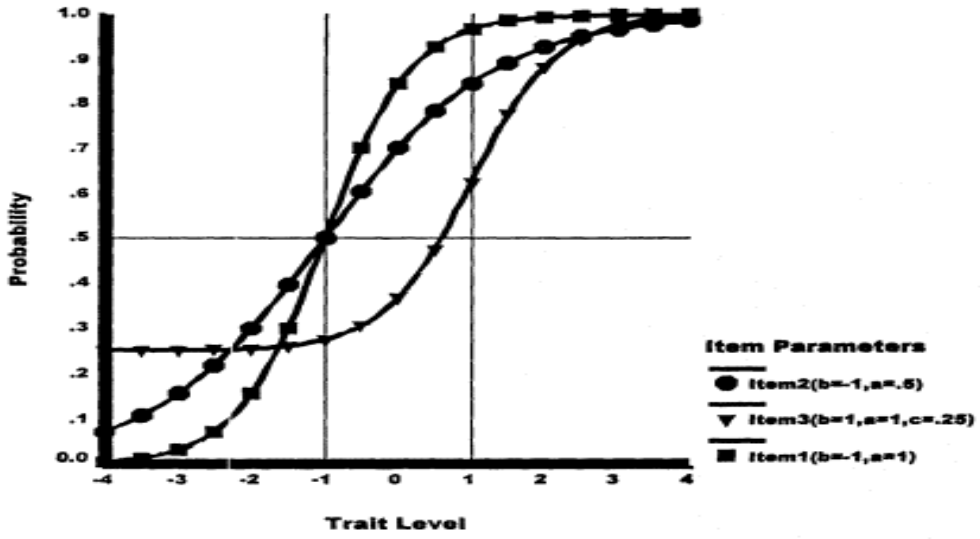

Figure 4. The ICCs for three items with different lower asymptote levels. Adapted from Item response theory for psychologists (p. 47), by Embretson, S., & Reise, S., 2000, Mahwah, NJ: L. Erlbaum Associates. Copyright 2000 by L. Erlbaum Associates.

The second assumption is that local independence of the item must be obtained. In other words, the responses to items are statistically independent in the test (Zhang, 2013).

This assumption is achieved if the relationships between items or persons are completely shaped by the IRT model. This means the probability of answering any item is independent of the outcome of any other item while controlling for person and item parameters (Embretson & Reise, 2000).

The conditional independence of items specifies the relationships between items in that there are no further relationships remaining after controlling model parameters. Moreover, the local independence condition is associated with the number of dimensions of the underlying latent traits; this is empirical evidence for unidimensionality. However, item local independence can be obtained for both unidimensional and multidimensional data if the IRT model includes person parameters for each dimension. Furthermore, Embretson and Reise (2000) indicated that item local independence can be achieved with more complex dependence items when items are linked or interactive with each other, which is accomplished by including the interaction effects or linkages in the model.

Dimensionality in IRT. IRT models have essentially been developed for instruments that are unidimensional. Unidimensionality suggests that items are measuring only one latent trait, and when the items measure more than one latent trait, the scale becomes multidimensional (Zhang, 2013). Unidimensionality is an important assumption for IRT models. At the same time, multidimensional IRT models have been developed for instruments that include more than one dimension. Instruments with multiple scales can be analyzed either concurrently using a multidimensional IRT model or consecutively using a unidimensional model. It was found that IRT models are

comparatively robust with moderate violations of strict unidimensionality (Harvey & Hammer, 1999).

IRT methods are typically used with instruments whose structure has already been determined using empirical methods such as exploratory or confirmatory factor analysis. Exploratory factor analysis can identify the dimensional structure of an instrument using data from a sample that represents the target population (Zhang, 2013). Next, confirmatory dimensionality analysis is used with another data set to confirm the number of dimensions (i.e. congruity between a known test structure and the statistical dimensional structure that is suggested by response data). Different dimensionality assessment procedures can be used, but all of them require similarities and distances between items.

There is a linear factor analysis (LFA) to assess dimensionality of item responses with linear relations. Nonlinear factor analysis can be alternatively used with both linear and nonlinear relations among items in a multidimensional scale. Another analysis that can be used to evaluate dimensionality is "dimensionality evaluation to enumerate contributing traits" (DETECT), which is based on item-pair conditional covariances; this method was modified in order to apply to test designs with unestimable item pairs such as occur with multistage design because of intended item nonresponse data (Zhang, 2013). These techniques can be used for either linear classical tests or for computerized adaptive tests.

For multistage testing, the different item responses collected at different stages are combined to get a final estimate of person proficiency, but the covariances of item pairs

cannot be estimated due to the multistage design. The similarities or distance of any two items cannot be estimated; this is called un-estimable item pair by design. In MST, there are no current methods that have the potential to be used with un-estimable item pairs to assess dimensionality. Recently, Zhang (2013) modified DETECT to apply to MST and any other test designs with un-estimable item pairs. However, this modification has not been thoroughly tested with real data.

IRT models are widely used in MST designs for item calibration and person ability estimation because IRT models have strong assumptions that ensure sample free person and item estimates, which is an important requerment for MST designs. Therefore, IRT is the major measurement model used for identifying item bias, equating, and establishing scoring rules. In this study, item parameters (item difficulty and item discrimination) of the item pool were needed for item assembly for each module (easy, medium, hard) in each condition, as well as for persons routing based on ability estimates in each module at each stage. Equating of different forms so that they are on the same scale is accomplished using IRT.

General Aptitude Test

The General Aptitude Test (GAT) is a large-scale assessment developed and administered by the National Center of Assessment in Higher Education in Riyadh, Saudi Arabia (S.A). The GAT is one of the requirements for colleges' and institutions' admission to higher education in Saudi Arabia for high school graduates. The purpose of the GAT is to predict the academic success of undergraduate students (Alanazi, 2014; Alshumrani, 2007; Dimitrov & Shamrani, 2015). There are two versions of the GAT, in

both the Arabic and English languages, which measure analytical and deductive skills of students who graduated from high school. According to the National Center for Assessment (NCA) in Higher Education (2017), the general focus of the GAT is to measure the students' capability for learning based on their proficiency in a particular area or topic. Five general abilities measured by the GAT are:

a) Reading comprehension;
b) identifying logical relationships;
c) resolving problems using basic mathematical concepts;
d) drawing inferences; and
e) measuring capacity.

The GAT is composed of two independent measures: verbal/qualitative (GAT-V) and math/quantitative (GAT-Q). The two sections include a varying number of items that assess examinee ability in different categories of content. The GAT-V includes items that measure four content areas, including:

a) Verbal reading comprehension (VRC), in which passages are given to the examinees and they are asked to answer comprehensive questions relating to the passages;
b) verbal analogy (VAN), in which the examinees are given a pair of words that they must match to the appropriate pair of words offered as multiple choices, by way of finding relationships between the two sets of pairs;

c) verbal contextual (VCA), in which the examinees are asked to find the synonym of a word from a list of choices that matches the meaning of the word given; and

d) verbal sentence completion (VSC); in which a short sentence is given to examinees, who are asked to fill in one or two blanks with missing words to complete meaningful sentences, drawing from the multiple choices given (Alanazi, 2014; NCA, 2017).

The GAT-Q focuses on mathematical problems that serve as examples of required skills related to problem-solving, logical reasoning, analysis, and measurements. The items in the GAT-Q measure the following five content areas:

a) Arithmetic (MAR),

b) Geometry (MGE),

c) Analysis (MAN),

d) Comprehension (MCO), and

e) Algebra (MAL) (Alanazi, 2014, NCA, 2017).

The duration of the GAT is two hours and a half. The test is divided into six parts, with 25 minutes allotted for each test part.

The GAT is a multiple-choice format test in which there exists only one correct answer per question; thus, the responses are scored as dichotomous (1=correct, 0=incorrect). The GAT is a 100-point test that carries a certain relative weight deciphered by the institutions to which examinees apply. According to the NCA (2017), the GAT scores are interpreted by relevant positioning of the examinees compared to other

students. Table 2 shows the scores and the related students' positions by percent performance bands.

Table 2.

The GAT Score and the relevant student's position as NCA clarified

Score	Student's position
81 and above	Top 5%
78 and above	Top 10%
73 and above	Top 20%
70 and above	Top 30%
65	The average
60 and below	Lowest 30%

The widespread use of the GAT in SA and some other Arabic countries, such as Egypt, Kuwait, Bahrain, and Oman, indicates the importance of investigating the reliability and validity of GAT scores. Thus far, small-scale studies have been conducted to investigate the validity of the GAT, with the intention of ensuring the accuracy of the admissions decisions. Alshumrani (2007) conducted a predictive validity study of GAT scores on first-year college GPA for 2,170 male students from Umm AlQura University, SA, controlling for high school GPA, high school major, and college major. The results indicated that the 8% of the variance in GPA was explained by the GAT. The reliability of the GAT-V was investigated by Dimitrov and Shamrani (2015) with a sample of 15,806 high school students from different regions of SA. The Cronbach's coefficient alpha (α) for internal consistency reliability of the test scores was 0.90 for all items (0.82 for Analogy, 0.75 for Sentence Completion, and 0.73 for Reading Comprehension).

Similarly, Alnahdi (2015) compared the predictive ability for three predictors: GAT score, high school GPA (HSGPA), and National Achievement Test score (NAT). This study was developed to predict the cumulative college GPA and graduation from the university using a sample of 27,420 students enrolled at Prince Sattam bin Abdulaziz University, SA. Alnahdi (2015) found that the GAT is a statistically significant predictor of cumulative college GPA and explains 12% of the GPA variation and 7% of the variation in student graduation status.

Unidimensionality is a significant assumption necessary in using item response theory (IRT) for test calibration. Thus, Dimitrov and Shamrani (2015) examined the dimensionality of the GAT-V and compared four different CFA models: (a) single factor model; (b) three-factor model, which uses three uncorrelated latent factors to show three different content specific domains; (c) three-factor model, which uses three correlated latent factors to show three content-specific domains; and (d) bifactor model, which uses a single general factor of verbal aptitude loading on all items and three latent factors as content-specific aspects of the general factor. The study concluded that the bifactor model that uses a single general factor and three content domains (Analogy, Sentence Completion, and Reading Comprehension) fit the data better than the other models. Therefore, the CFA modeling of GAT-V data suggests that the GAT is fundamentally unidimensional, which justifies the use of IRT calibration of items and scoring of performance.

In order to know whether GAT items exhibit bias across various subgroups, Sideridis and Tsaousis (2013) used four different statistical strategies: uniform and non-

uniform differential item functioning (DIF); measurement invariance for multi-group confirmatory factor analysis; and differential test functioning (DTF), using a large sample size (41,134 examinees). The different methodologies were applied across gender (male vs. female), provinces (Saudi Arabia vs. Bahrain), school type (public vs. private), and thirteen different regions, using five forms (A to E). The technical report suggested that the GAT items did not exhibit obvious uniform or non-uniform DIF and the effect size (Cohen's *d*) was smaller than the threshold of 0.35; this was the case for all but two items, which showed noticeable non-uniform DIF. For measurement invariance of the GAT, the results validate the presence of configural and metric invariance across all population comparisons. The conclusion of GAT' results indicate that the GAT possessed negligible amounts of bias across gender, province, school type, test form and region. Therefore, GAT is stable regarding its function for all factors tested (Sideridis & Tsaousis, 2013).

Given important utilization and stability of the GAT, more restrictive analytical studies are strongly needed to enhance the validity of the GAT's applications. Further studies would be additionally helpful to examine the GAT based on demographic information of examinees, schools, curricula, and regions in order to understand the psychometric features of the General Aptitude Test.

Chapter Two: Method

Procedure

The GAT is usually administered in paper format twice during a school year (Fall, Spring). Therefore, each examinee has two opportunities to take the GAT paper test per academic year on scheduled dates. The examinees can also take the computer-based test at various times during the academic year. The GAT is a timed test; the duration of the test is two and half hours. The test is divided into six parts. Examinees have 25 minutes to answer and review the questions on each part. Then, they must move to the next part, so they are not allowed to review the previous parts when the new part is presented (NCA, 2017)

The test was administered, and data were collected by the National Center of Assessment (NCA) in Higher Education, Riyadh, Saudi Arabia. The author obtained secondary data from the NCA, as well as the agreement to use the data for research purposes. Data were provided to the researcher via an emailed file. The data are a frame that represents the population of examinees who annually take the GAT. The data were anonymous and included responses to all items with no missing data, and only demographic information about sex and region were available.

For the current study, data from the GAT paper-based administration were used. The data comprised 45,738 examinees that took the General Aptitude Test in Fall 2015, administrated in 13 different regions. In order to have a suitable sample, 20% of the

dataset was selected for a random sample; precisely, 9,108 examinees, 4,183 males (45.9%) and 4,925 females (54.1%).

Instrument

The instrument used to create the multiple forms of MST is the GAT developed by the NCA. According to the NCA (2017), the GAT aims to measure analytical and deductive skills of students (more information about GAT was presented in the literature review. The GAT comprises two independent measures:

(a) Section1, measuring verbal ability (GAT-V,) and

(b) Section 2, measuring math or quantitative ability (GAT-M).

The GAT consists of 96 items divided into 52 items in the verbal section and 44 items in the quantitative section. Each section has several domains; Table 3 lists these domains and the number of items in each section. The items are in multiple-choice format where there is only one correct response. Thus, the GAT data are dichotomous with a score of "one" for a correct response and "zero" for each incorrect response.

Table 3.

The Number of Items in each Domain of the Verbal and Quantitative/Math GAT

GAT-V	Number of items	GAT-M	Number of items
Reading comprehension (VRC)	20	Arithmetic (MAR)	16
Analogy (VAN)	16	Geometry (MGE)	8
Contextual (VCA)	10	Analysis (MAN)	8
Sentence completion (VSC)	6	Comprehension (MCO)	8
		Algebra (MAL)	4
Total	52		44
Total after deletion	48		43

Assessment of the GAT

The CLT was the item pool for MST designs. The initial item analysis using CFA and 2PL IRT was conducted. Four items (VSC-D, VSC-E, VRC, VRC-S) in GAT-V and one item (MAR-L) in GAT-M were deleted because their discrimination indices were negative. The CLT version of GAT subtests was reassessed after deleting weak items from both sections. The unidimensionality of the GAT was separately assessed for GAT-V and GAT-M via CFA using SPSS Statistic Version 25 and AMOS Version 22. The CFA analysis supported the GAT-V and GAT-M as unidimensional tests; the existence of one factor was not rejected. See Appendix C for GAT-V and Appendix D for GAT-M. The parameters of the item pool were estimated by IRT-2PL using BILOG-MG3. Tables 1 and 2 in Appendix A display the values of the parameters of the item pool for each section.

The calibration for the GAT-V showed that the range of the item difficulty parameter (d) was from -2.59 to 4.66 with a mean of 0.232 and standard deviation (*SD*) of 1.51. The mean of the discrimination parameter (a) was 0.45 with *SD* of 0.20, ranging from 0.079 to 0.98. For GAT-M, the item difficulty parameters ranged from -1.88 to 2.91 with a mean of 0.324 and *SD* of 0.96. The discrimination parameter (a) ranged from 0.079 to 0.980 and the mean was 0.465 with *SD* of 0.16. The reliability indices were 0.86 for GAT-V and 0.85 for GAT-M.

The tenability of the functional form was assessed by examining the empirical test information function curve (TIC) and item characteristic curves (ICCs). The TIC of GAT-V represented in Figure 1 (solid line) in Appendix B that shows that the test information covered ability ranging from -4 to +4 with the peak information found when the theta level was -0.50; the curve was a little skewed toward the low level of ability more than the high ability levels, indicating that the information did not cover well the whole range of ability levels. The highest information coverage of ability centered between logit ability -2 and +1. The figure also shows that the standard error curve (dotted line) was inversely related to the test information curve. The standard errors (*SE*) were reduced at the low and medium levels between -2 and +1 of theta while it increased above +1 and below -2 level of ability. Moreover, the power of discrimination was somewhat higher for examinees between these two levels of theta. Additionally, there was little information about those who had ability levels at both edges: the test information was somewhat poor for ability levels higher than + 2 and lower than -2. In sum, this indicates the acuracy of ability estimation was not equal along the continuum of

the ability scale which was expected because fewer items and so less information was available at the upper and lower edges of the ability scale.

The TIC of GAT-M represented in Figure 2 (solid line) in Appendix B shows that the test information covers ability ranging from -4 to +4. The peak information was in the center of theta at a level of zero, the curve was normal shape, so the information was covered well between -2 and +2. Consequently, the standard error curve (dotted line) decreased at the average ability levels between -2 and +2 of the theta continuum. The *SE* increased above the +2 and below -2 level of ability, implying the acuracy of ability estimations was not equal along the continuum of the ability scale because the test information was considerably less for ability levels higher than + 2 and lower than -2.

MST Panel Structure Forms

The two subtests, verbal and quantitative, are assumed to be independent and they measure different content; thus, the multistage tests were designed separately, one for GAT-V and another for GAT-Q/M. MST panels were assembled from an item pool that contained 48 items for GAT-V and 43 items for GAT-M. The next section explains the construction of the MST panels that were applied with both sections of the GAT-V and GAT-M. Two panel structures with two different test lengths and two assembly conditions were constructed.

The two-stage MST panel structure designs. The first panel design included two stages, 1-3 panel structures that had one module for the first stage (the routing stage) with nine items, and three modules for the second stage (easy, medium, and hard

modules) with three items at each module; the test length was twelve items. See Figure 5. For the two-stage MST, there were three potential paths:

a) Low-path that includes module 1M (routing) and module 2E (easy),
b) Middle-path that includes module 1M (routing) and module 2M (medium), and
c) High-path that includes 1M (routing) and module 2H (hard).

The three-stage MST panel structure designs. The second panel design included three stages, 1-3-3 panel structures that had one module for the first stage (the routing stage) with nine items, and three modules for the second and third stage (i.e., easy, medium, and hard modules.) There were three items at each module, which makes the test length 15 items. For the three-stage MST, there were nine potential paths because the nonadjacent routing was used (more details about nonadjacent routing can be found in routing method section). The nine paths were:

a) Low-path includes module 1M (routing), 2E and 3E (EE);
b) Low-medium-path includes 1M (routing), 2E and 3M (EM);
c) Low-high-path includes 1M (routing), 2M and 3H (MH);
d) Middle-low-path includes module 1M (routing), 2M and 3E (ME);
e) Middle-path includes module 1M (routing), 2M and 3M (MM);
f) Middle-high-path includes module 1M (routing), 2M and 3H (MH);
g) High-low-path includes 1M (routing), 2H (hard), and 3E (HE);
h) High-medium-path includes 1M (routing), 2H (hard), and 3M (HM); and
i) High-path includes 1M (routing), module 2H (hard), and 3H (hard).

Assembly Conditions

Figure 5 shows all designs for two- and three-stage MST panel structures. All MST designs had two different assembly conditions for the second or third stage(s). The first assembly condition was the small-difference or narrow range (NR) condition, where the three modules at the second stage (for 1-2 and 1-3-3) and at the third stage (for 1-3-3 MST) included items with slight differences in difficulty from each other; thus, the range in item difficulty was narrow.

In contrast, in the large-difference or wide range (WR) condition, the three modules at the second stage (for 1-2 and 1-3-3) and at the third stage (for 1-3-3 MST) included items with significant differences from each other; thus, the range was wide. For all panel designs, the test information function (TIF) was fixed at 0.0 for the routing stage and medium modules, one standard deviation below the average of the item difficulty for easy modules, and one standard deviation above the average of the item difficulty for hard modules on the continuum "theta."

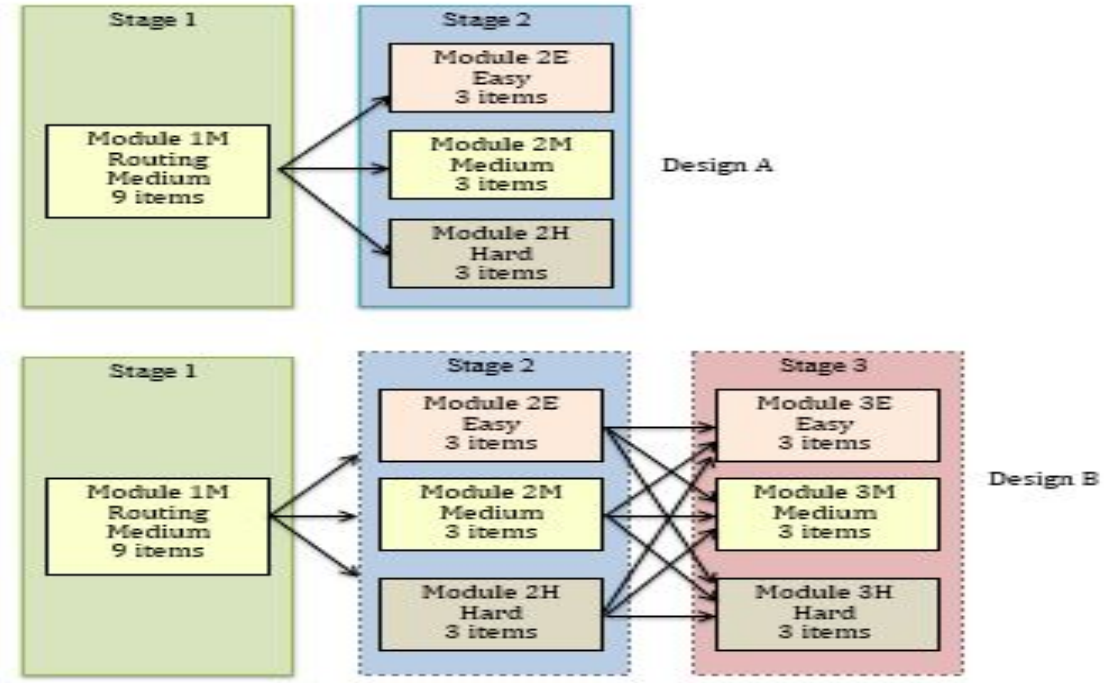

Figure 5. Two- and Three-Stage MST panel structure designs

Routing Methods

Two routing methods were used to assign examinees to the next stage(s). The methods were Defined Population Intervals (DPI), and the Approximate Maximum Information method (AMI). The two approaches were described in the literature review section. In the MST design that used DPI, the examinees were sorted after each stage in rank order according to their score; then they were assigned to the next module using a set percentage to each module. The percentages 30%, 40%, and 30% were used to assign examinees to the easy, medium, and hard modules, respectively.

The second MST design used the AMI method. The examinees were routed into the next stage using the test information functions (TIFs); the connection between the cumulative TIFs of adjacent modules was determined. The cutscores of estimated ability

-0.524 and 0.524 were used to assign examinees to the low class (below average), medium class (average), and high class (above average).

Nonadjacent routing was used in this study for equity purposes. In MST with more than two stages, there are two strategies of routing (adjacent vs. nonadjacent) regarding allowing examinees to be routed into all avaliable modules in the next stages. In practiclity, adjacent routing is usually used in MST studies where examinees can only be routted to modules in the next stage that are adjacent in level of difficluty related to thier module in the current stage (Kim & kim, 2018). For example, in 1-3-3 design, if the adjacent routing is used, examinees who are assigned to the easy module in the second stage, have only a chance to be assigned to the easy or medium module, not to the hard. They cannot recover from low performance on stage 1 and cannot be aasigned to the hard module in the thrid stage, even if their performance is high in the easy module at the second stage. However, from the perspective of fairness and equity, the nonadjacent routing allows examinees to recover from the bad performance in the first stage, so they can be routed to the hard module if their performance is high enough in the easy module at second stage. In short, nonadjacant routing allows to have more paths than adjacant routing; for example, in a 1-3-3 design, there are nine possible paths if nonadjacent routing is used whereas there are only seven paths if adjacent routing is used

Data Analysis

Descriptive item parameters (difficulty, discrimination) were obtained to generate the MST panel structures. The mean and *SD* of item difficulty were used to divide the items into three categories. The items with difficulty levels (d) that were above one *SD* of

the mean were assigned to the hard modules; the items with difficulty levels below one *SD* of the difficulty mean were assigned to easy modules. The rest of the items were assigned to medium modules. See Table 1 and Table 2 in Appendix A for more detail. Additionally, an initial analysis was conducted to examine the dimensionality of each subtest (GAT-V and GAT-M). To answer the research questions, different analyses and indices were used. BILOG-MG3 was used to estimate ability estimates and the standard errors using IRT with a 2PL model. The next section provides a description of the analysis for each question, separately.

Research question 1. Does the performance of two assembly conditions (NR, WR) yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V/ GAT-M?

This question has two parts regarding comparison of the results of two assembly conditions in terms of ability estimates and accuracy in classification decisions. For ability estimation, the 2PL IRT model was run for all conditions to obtain the ability estimates and the standard errors (*SE*) for estimated ability to compare; the assembly condition with the smaller average *SE* was considered to be more accurate.

For classification decisions each examinee either in CLT or MST was placed into one of three classes using DPI and AMI rules. For DPI conditions (MST and CLT), the top 30% of examinees were assigned to the high class (above average), 40% of them to the medium class (average), and the lowest 30% to the low class (below average). Regarding AMI conditions (MST and CLT), the examinee whose ability was 0.524 and higher was assigned to the high class (above average), the examinee whose ability was -

0.524 or less was assigned to the low class (below average), and the rest of the examinees, having had an estimated ability between -0.524 and 0.524, were assigned to the medium class.

If both the ability scores on CLT and the estimated abilities on MST equally classified examinees as above average, average, and below average, this means the correct decision was made. If the estimated ability on CLT and on MST differently classified the examinees, this means the wrong decision was made. The percentage of agreement in classification between CLT and MST was considered as a percentage of correct classification, where the CLT was considered as the correct classification. The assembly condition that yielded a higher count of correct decisions distinguished it as the more accurate model. The percentage of classification errors was calculated. Therefore, all conditions were compared in terms of correct classification rate and false error rate (either false negative or false positive error.) Thus, the assembly condition with the highest percentage of correct classification was deemed the more accurate condition. All analysis procedures were independently applied for both GAT-V and GAT-M.

Research question 2. Does the performance of DPI and AMI routing methods yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V and GAT-M?

To answer this question, a comparison between models that used the DPI method and models that used the AMI method was conducted under all conditions. The models were compared in terms of the amount of the error associated with ability estimates; models with a smaller average error were considered to be the more precise models. As in

research question 1, the MST model that yielded the highest proportion of correct classification decisions was deemed the most precise model. All analysis procedures were independently applied for both GAT-V and GAT-M.

Research question 3. Does the performance of two-stage MST and three-stage MST yield comparable results in terms of ability estimates and accuracy in classification decisions for both GAT-V and GAT-M?

As in research questions 1 and 2, the analysis for the third research question was conducted in the same manner but compared the performance of 1-3 MST and 1-3-3 MST. The model with the smallest average error of ability estimates was the most precise model and the model with the highest percentage of correct classification decisions was the most accurate model.

Research question 4. Do the examinees' ability scores obtained from classical testing correlate positively and statistically significantly to the examinees' scores on the MST for each condition for both GAT-V and GAT-M?

For all conditions and for both sections of the GAT, the bivariate correlation between students' scores and the students' positions (using IRT calibration) on each MST condition were calculated. The Pearson correlation coefficient was used as index for the relationship. Coefficients of 0.70 and above were considered as strong correlations as well as a strong index of adequacy for MST as an alternative to classical linear testing. Table 4 provides a descriptive of the aim of each research question and the target outcomes that were used to evaluate the performance of each condition in MST for GAT.

Table 4.

Brief Description of Conditions and Outcome Measures

Target Point	Conditions	Outcome Measures
Q1: Assembly conditions	Two conditions: - small-difference or narrow range (NR) - large-difference or wide range WR	- Ability estimation and Standard errors: model with lower average *SE* is more accurate. - Classification decisions: Model with highest percentage of correct classification decisions between examinees classes on CLT and estimated abilities classes on MST is more accurate.
Q2: Routing methods	Two conditions: - Defined Population Intervals (DPI) - Approximate Maximum Information (AMI).	- Ability estimation and Standard errors: models with a smaller average error are the more precise models. - The MST model that yields the highest proportion of correct classification decisions is the most precise model
Q3: Number of stages	- Two conditions: - Two stages (1-3) - Three stages (1-3-3)	- Ability estimation and Standard errors: models with smaller average error are the more precise models. - The MST model that yields the highest proportion of correct classification is the most precise model
Q4: Relationship to Raw Score	All previous conditions	- Bivariate correlation between examinees' ability scores on CLT and the students' positions on MST - Coefficient of 0.70 and above is considered a strong correlation.

Chapter Three: Results

Presented in this chapter are the results of the series of analyses, the purpose of which was to assess the multiple form designs of multistage testing for GAT. The research questions are addressed in analyses of results of sequential MST designs using IRT. BILOG-MG3 was used for IRT analysis with a 2PL for assessing MST designs while SPSS Version 25 was used for statistical analyses, such as descriptive, *t*-tests, correlations, percentages, and classification.

Research Question One

Does the performance of two assembly conditions yield comparable results in terms of ability estimation and accuracy in classification decisions for GAT-V and GAT-M?

The two assembly conditions of items, narrow range (NR) and wide range (WR) of item difficulty, were applied to the second stage in MST-2stage, and to the second and third stage in MST-3stage for both subtests of the GAT. The comparison was conducted among all conditions in terms of ability estimates and accuracy in classification decisions. The results are presented separately in two sections: GAT-V and GAT-M.

GAT-V: Comparison between NR and WR in estimated ability. A comparison between NR and WR conditions in accuracy of estimated ability for GAT-V was conducted separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-V with MST-2stage. The examinees' abilities were estimated in the routing stage, which included nine items, and in the second stage, which included three items in each module for both conditions NR and WR. Table 5 shows the descriptive statistics of item difficulty, aggregated over items in each module for the MST-2Stage GAT-V. It is notable that in the second stage the item difficulty range was 3.6 in the NR condition while it was nearly twice that (7.3) in the WR condition. The ability means and error means were estimated in NR and WR for all routing method conditions.

Table 5.

Descriptive Statistics of Item Difficulty in Each Module in MST-2Stage GAT-V

Difficulty difference	Stage	Module/ level	# items	M	SD	Range	Item ID
NR	1	Routing	9	-0.11	0.75	2.2	VAN-A, VAN-J, VAN-N, VCA-B, VCA-H, VSC VRC-K, VRC-M, VRC-O
	2	Hard	3	0.34	1.55	3.6	VSC-A, VRC-C, VRC-Q
	2	Medium	3				VAN-D, VCA-G, VRC-D
	2	Easy	3				VAN-H, VCA, VRC-E
WR	1	Routing	9	-0.11	0.75	2.2	VAN-A, VAN-J, VAN-N, VCA-B, VCA-H, VSC VRC-K, VRC-M, VRC-O
	2	Hard	3	0.68	2.83	7.3	VAN-F, VRC-H, VRC-L
	2	Medium	3				VAN-D, VCA-G, VRC-D
	2	Easy	3				VAN-I, VRC-B, VRC-N

Results indicate there was a minuscule difference between NR and WR regarding the targeted estimates in all conditions. The ability means were quite similar in NR (0.892 and 0.998) and WR (0.882 and 0.906) in DPI and AMI conditions, respectively. Although the differences among ability means were small, these differences were significant at $p < 0.001$ with *df* of 9,107. The results of *t*-tests are provided in Table 6, which reveals that

the mean ability in NR was significantly higher than the mean ability in WR for both routing conditions DPI and AMI.

According to Table 6, the standard error (*SE*) means for estimated ability were also similar in all conditions, resulting in 0.899 and 0.910 in NR and 0.916 and 0.924 in WR. The *t*-test results show that the NR *SE* means (0.899 and 0.910) were statistically significantly lower than the *SE* means (0.916 and 0.924) of WR conditions, for both DPI and AMI conditions.

Table 6.

Differences between NR and WR among All Conditions for MST-2Stage of GAT-V

Condition	Ability Mean		*t*-test	SE Mean		*t*-test
	NR	WR		NR	WR	
DPI	0.982	0.882	112.12**	0.899	0.916	1254.4**
AMI	0.998	0.906	118.48**	0.910	0.924	1362.2**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method,
** $p < 0.001$ with *df* 9,107

The boxplots for ability means in all conditions for MST-2Stage are displayed in Figure 6. The top figures represent NR conditions and the bottom figures correspond to WR conditions. For AMI conditions (left side of the figure), the groups' mean ability estimates of GAT-V in NR conditions were higher than the groups' means in WR conditions, where outliers existed in group 2 (medium group) of WR conditions; this may have pulled the mean down. The ability means of the NR were slightly higher than the means in the WR conditions. For DPI conditions (right side of the figure), the groups' means were quite similar between the two conditions. Group 1 (low group) had similar mean in both conditions NR and WR, but the WR had outliers in group 1. The outliers generally impact the estimations, so they are smaller or larger than the central value,

depending on the direction of the outliers. The outliers were in the low band in group 1 for the WR condition which made the mean smaller than it likely should be. Thus, the mean of this group was underestimated because of the presence of outliers.

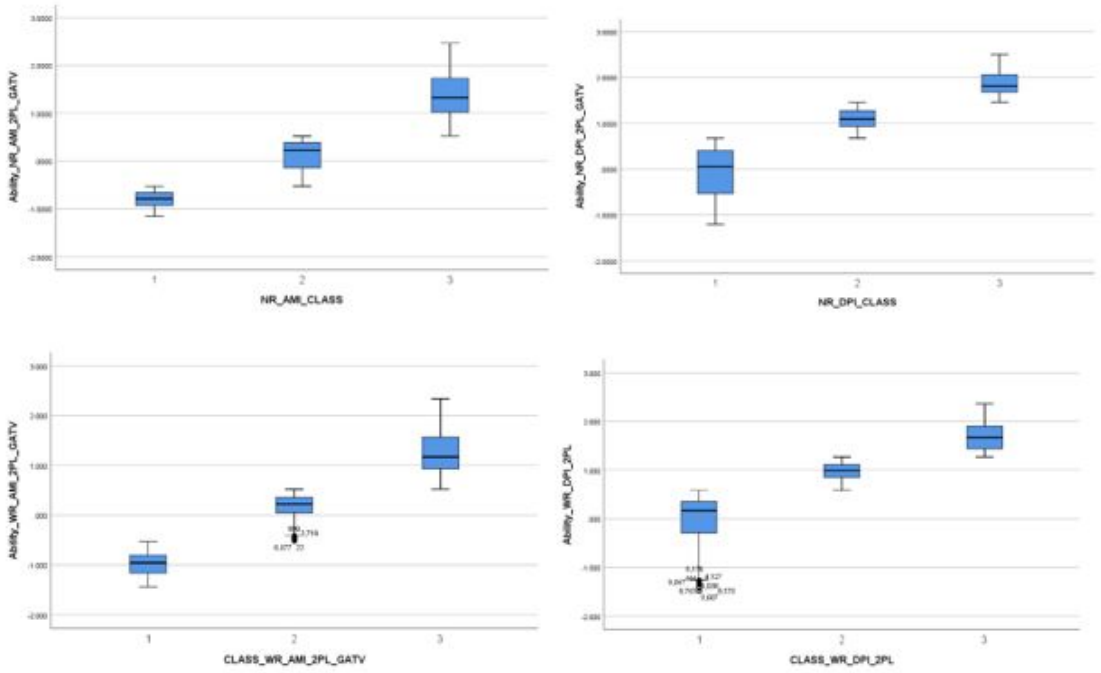

Figure 6. Classes in all conditions in MST-2Stage for GAT-V, NR-AMI and WR-AMI (left), NR-DPI and WR-DPI (right)

GAT-V with MST-3Stage. The MST-3Stage consisted of three stages. The routing stage included nine items, while the second and third stages included three items in each module for both conditions, NRWR (NR at stage 2 and WR at stage 3), and WRNR (WR at stage 2 and NR at stage 3). The item summary information for each module in MST-3Stage of GAT-V is provided in Table 7.

Table 7.

Descriptive Statistics of Item Difficulty in Each Module in MST-3Stage GAT-V

Difficulty difference	Stage	Module/ level	# items	M	SD	Range	Item ID
NR at Stage 2 & WR at Stage 3	1	Routing	9	-0.11	0.75	2.22	VAN-A, VAN-J, VAN-N, VCA-B, VCA-H, VSC VRC-K, VRC-M, VRC-O
	2	Hard	3	0.34	1.55	3.60	VSC-A, VRC-C, VRC-Q
	2	Medium	3				VAN-D, VCA-G, VRC-D
	2	Easy	3				VAN-H, VCA, VRC-E
	3	Hard	3	0.78	2.81	7.25	VAN-F, VRC-H, VRC-L
	3	Medium	3				VAN-G, VCA-F, VRC-A
	3	Easy	3				VAN-I, VRC-B, VRC-N
WR at Stage 2 & NR at Stage 3	1	Routing	9	-0.11	0.75	2.22	VAN-A, VAN-J, VAN-N, VCA-B, VCA-H, VSC VRC-K, VRC-M, VRC-O
	2	Hard	3	0.68	2.83	7.26	VAN-F, VRC-H, VRC-L
	2	Medium	3				VAN-D, VCA-G, VRC-D
	2	Easy	3				VAN-I, VRC-B, VRC-N
	3	Hard	3	0.44	1.53	3.36	VSC-A, VRC-C, VRC-Q
	3	Medium	3				VAN-G, VCA-F, VRC-A
	3	Easy	3				VAN-H, VCA, VRC-E

The means of estimated abilities were slightly different between NRWR (1.298 and 1.246) and WRNR (1.218 and 0.935) using DPI and AMI conditions, respectively. The *t*-test results are presented in Table 8 indicated there was a significant difference

between NRWR and WRNR in the mean of estimated abilities. The ability means were higher in NRWR conditions than the means in WRNR conditions, using the same routing method, DPI or AMI.

Table 8 shows there was a small difference between the standard error means of estimated abilities in NRWR (0.862 and 0.940) and WRNR (0.882 and 0.935), either in DPI or AMI conditions. Nevertheless, the *t*-test results showed these small differences were statistically significant at $p < 0.001$ with degrees of freedom of 9, 107. For the DPI method, the NRWR condition had a significantly smaller *SE* mean than the *SE* mean of WRNR, while the WRNR had a significantly smaller *SE* mean than the NRWR condition with the AMI method, see Table 8 for details.

Table 8.

Differences between NR and WR among All Conditions for MST-3Stage of GAT-V

Condition	Ability Mean		*t*-test	*SE* Mean		*t*-test
	NRWR	WRNR		NRWR	WRNR	
DPI	1.298	1.218	108.5**	0.862	0.882	1258.4**
AMI	1.246	0.962	163.2**	0.940	0.935	1217.4**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with *df* 9,107

The classes' means of NRWR conditions (top figures) were slightly higher than in the WRNR conditions (bottom figures), in which the same routing method was used, either DPI or AMI (Figure 7). The differences were noticeable in class 2 and 3 with DPI conditions. Group 1 in NRWR-DPI and WRNR-DPI were quite similar; the existence of outliers in class 1 may have pulled the mean slightly down in the WRNR-DPI condition. For AMI conditions, the top edge of the boxplot for class 3 in NRWR-AMI reached an

estimated ability about 3.5 while the class 3 in WRNR-AMI reached ability level of ~ 2.5, which made the mean slightly higher in this group in NRWR-AMI compared to class 3 in WRNR-AMI. There were outliers in class 1 and class 3 in the WRNR-AMI condition, which dropped the mean lower relative to the means of the same classes in the NRWR-AMI condition.

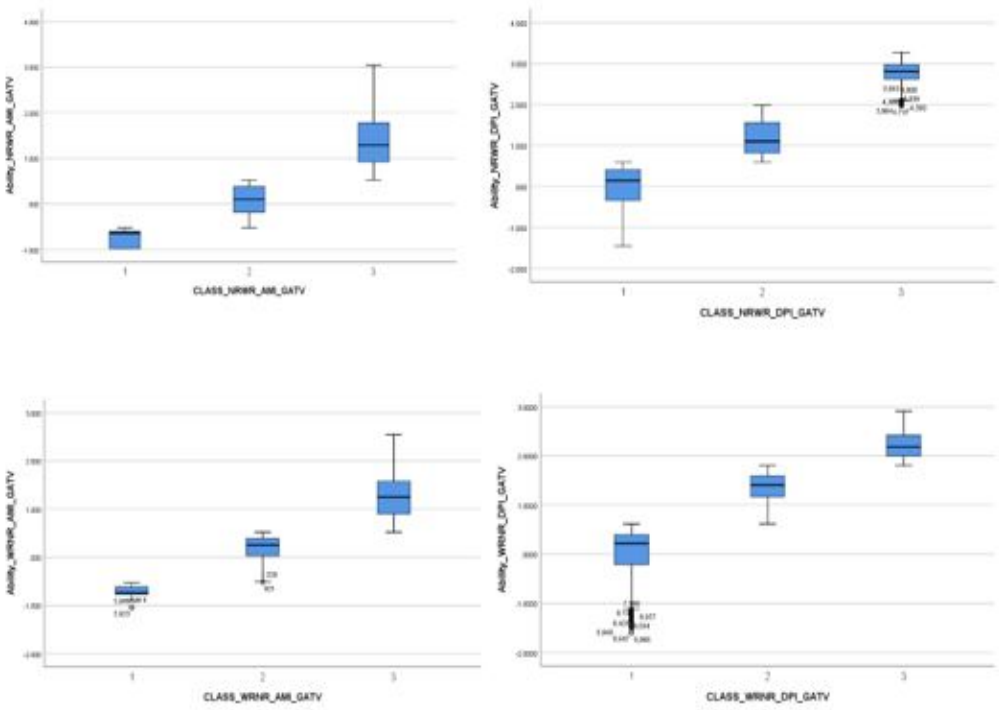

Figure 7. Classes in all conditions in MST-3Stage for GAT-V, NRWR-AMI and WRNR-AMI (left), NRWR-DPI and WRNR-DPI (right)

GAT-V: Comparison between NR and WR in accuracy of classification decisions in MST. The examinees were classified into three classes/levels according to their estimated abilities in MST GAT-V. For DPI conditions, the top 30% of examinees were assigned to the high class (above average), 40% of them to the medium class (average), and the lowest 30% to the low class (below average). Regarding the AMI

conditions, the examinee whose ability was 0.524 and higher was assigned to the high class (above average), the examinee whose ability was - 0.524 or less was assigned into the low class (below average), and the rest of the examinees, having had an estimated ability between -0.524 and 0.524 and were assigned to the medium class.

Table 9 displays the percentages of correct classification, false positive, and false negative decisions. The "correct" decision was made when the examinees were classified into the same class in MST and CLT. However, if the examinees were classified into classes in MST higher than their classes in CLT, a false positive (FP) classification decision was made. Lastly, the false negative (FN) classification decision was made when the examinees were classified into a lower class in MST than their class in CLT.

Table 9.

Percentage of Correct Classification Decisions in All Conditions of MST GAT-V

Condition	Precision of classification decision	MST-2STAGE NR %	MST-2STAGE WR %	MST-3STAGE NRWR %	MST-3STAGE WRNR %
DPI	CD	57.9	53.5	63.4	60.4
	FP	20.4	22.3	17.6	20.4
	FN	21.7	24.2	19.0	19.3
AMI	CD	39.6	39.1	31.7	35.7
	FP	60.4	60.7	68.3	64.3
	FN	0.0	0.2	0.0	0.0

CD=correct decision, FP= false positive decision, FN= false negative decision

The percentage of correct classification decisions were quite similar in the NR and WR conditions in which the same routing method was used, ranging from 57.9 % to 39.1% as presented in Table 9. For MST-2Stage, the percentage of correct classification decisions was slightly higher in NR conditions (57.9% and 39.6%) than in WR conditions (53.5% and 39.1%), for both DPI and AMI conditions, respectively. It is noticeable that

NR and WR within the AMI condition *incorrectly* classified more than 60% of examinees into classes higher than their classes in CLT, which presented a case of a majority of false positive decisions.

For MST-3Stage, the NRWR with the DPI condition showed that the highest percentage of correct classification decisions was 63.4% with 36.6% incorrect decisions. Of the incorrect decisions, 19.0% were false negative decisions, and 17.6 % were false positive decisions. However, the WRNR showed a higher percentage of correct classification decisions (35.7%) when compared to the NRWR (31.6%) in the AMI condition. The highest percentages of incorrect decisions were false positive classification decisions (68.3% and 64.3) in NRWR and WRNR, respectively, wherein the examinees of the MST-3Stage were incorrectly assigned to classes that were higher than their classes in CLT.

GAT-M: Comparison between NR and WR in estimated ability. A comparison between NR and WR conditions in accuracy of estimated ability for GAT-M was conducted separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-M with MST-2Stage. The examinees' abilities were estimated in the routing stage, which was comprised of nine items, as well as in the second stage, which included three items in each module for both WR and NR conditions. Table 10 shows the descriptive statistics of item difficulty, aggregated over items, in each module for MST-2Stage of GAT-M. The item difficulty range was 2.78 in the second stage of the NR condition, while it was 4.79 in WR, which was nearly double that of NR.

Table 10.

Descriptive Statistics of Item Difficulty in Each Module in MST-2Stage GAT-M

Difficulty difference	Stage	Module/ level	# items	M	SD	Range	Item ID
Small/ NR	1	Routing	9	0.44	0.48	1.43	MAR-C, MAR-F, MAR-J, MGE-B, MGE-D, MAL-A, MAL-B, MCO-B, MCO-F
	2	Hard	3	0.23	1.09	2.78	MAR-A, MRR-O, MGE-E
	2	Medium	3				MAR-H, MGE-F, MAN-A
	2	Easy	3				MGE, MAN-C, MCO
Large/WR	1	Routing	9	0.44	0.48	1.43	MAR-C, MAR-F, MAR-J, MGE-B, MGE-D, MAL-A, MAL-B, MCO-B, MCO-F
	2	Hard	3	0.37	1.66	4.79	MAR-G, MAN, MCO-G
	2	Medium	3				MAR-H, MGE-F, MAN-A
	2	Easy	3				MAR, MGE-A, MAN-D

The differences in ability means in NR and WR were small; the means of abilities were slightly larger in the WR (0.87 and 0.89) than the ability mean in the NR (0.85 and 0.87) conditions for both DPI and AMI (see Table 11). The results of *t*-tests revealed significant differences between the WR and NR regarding means of estimated abilities. The mean in WR conditions was significantly higher than the mean in NR conditions in which the same routing method was utilized, either DPI or AMI.

The *SE* means were quite similar in NR (0.95) and in WR (0.94) for all conditions. However, these small differences were statistically significant at $p < 0.001$ with *df* of 9,107, meaning the standard errors were smaller in WR than in NR, as seen in Table 11.

Table 11.

Differences between NR and WR Among All Conditions for MST-2Satge for GAT-M

Condition	Ability Mean		*t*-test	*SE* Mean		*t*-test
	NR	WR		NR	WR	
DPI	0.851	0.870	103.96**	0.950	0.939	2138.8**
AMI	0.877	0.895	102.55**	0.947	0.935	2100.3**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with *df* 9,107

Figure 8 illustrates the degree to which the groups differed; the groups' means were similar in NR and WR conditions. The figures additionally show outliers in the NR-DPI (top right), which make the mean higher than the mean in the NR-AMI condition. However, the mean of all groups in NR conditions (top figures) were generally like the groups' means in WR conditions (bottom figures), while the range of class 3 (high group) in NR and WR conditions with AMI were wider than in DPI conditions.

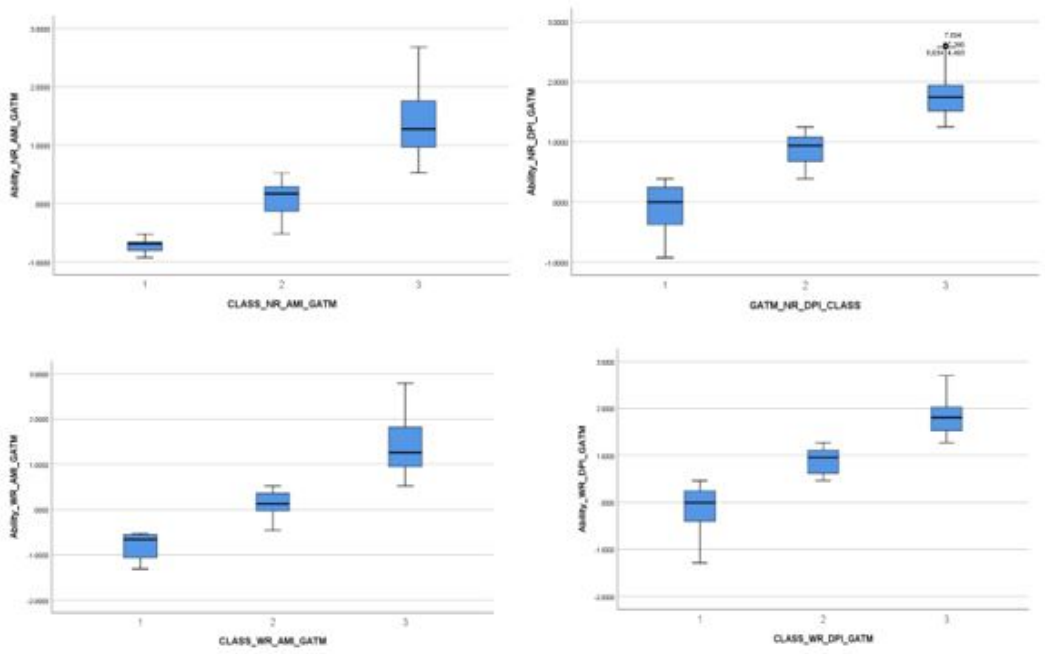

Figure 8. Classes in all conditions in MST-2Stage for GAT-M, NR-AMI, and WR-AMI (left), NR-DPI and WR-DPI (right)

GAT-M with MST-3Stage. The examinees' abilities were estimated in the routing stage (nine items), second stage (three items per module), and in the third stage (three items per module), for both conditions: NRWR condition (NR at stage 2 with WR at stage 3) and WRNR condition (WR at stage 2 with NR at stage 3). Table 12 shows the aggregate item information for each module in the MST-3Stage of GAT-V.

Table 12.

Descriptive Statistics of Item Difficulty for Each Module in MST-3Stage GAT-M

Difficulty difference	Stage	Module/ level	# items	M	SD	Range	Item ID
NR at Stage 2 & WR at Stage 3	1	Routing	9	0.44	0.48	1.43	MAR-C, MAR-F, MAR-J, MGE-B, MGE-D, MAL-A, MAL-B, MCO-B, MCO-F
	2	Hard	3	0.23	1.09	2.78	MAR-A, MRR-O, MGE-E
	2	Medium	3				MAR-H, MGE-F, MAN-A
	2	Easy	3				MGE, MAN-C, MCO
	3	Hard	3	0.37	1.66	4.79	MAR-G, MAN, MCO-G
	3	Medium	3				MAR-H, MGE-F, MAN-A
	3	Easy	3				MAR, MGE-A, MAN-D
WR at Stage 2 & NR at Stage 3	1	Routing	9	0.44	0.48	1.43	MAR-C, MAR-F, MAR-J, MGE-B, MGE-D, MAL-A, MAL-B, MCO-B, MCO-F
	2	Hard	3	0.37	1.66	4.79	MAR-G, MAN, MCO-G
	2	Medium	3				MAR-H, MGE-F, MAN-A
	2	Easy	3				MAR, MGE-A, MAN-D
	3	Hard	3	0.23	1.09	2.78	MAR-A, MRR-O, MGE-E
	3	Medium	3				MAR-H, MGE-F, MAN-A
	3	Easy	3				MGE, MAN-C, MCO

The means of estimated abilities for examinees were slightly different in all conditions; WRNR had higher ability means (0.99 and 0.76), compared to the ability

means for NRWR conditions (0.76 and 0.47) using the same routing method, either DPI or AMI. Moreover, the *t*-test results indicated there was a statistically significant difference between NRWR and WRNR in the mean of estimated abilities at $p < 0.001$; the ability means were lower in the NRWR than in the WRNR conditions. See Table 13.

Additionally, there were slight differences between the *SE* means for estimated abilities in NRWR (0.98 and 1.00) and WRNR (0.94 and 0.97) using DPI and AMI conditions, respectively. However, the *t*-test results, are presented at Table 13, indicated that these slight differences were statistically significant at $p < 0.001$ with $df = 9,107$. The WRNR conditions had significantly smaller standard errors of estimated ability than the NRWR, which means the estimated ability was more accurate in WRNR conditions than in NRWR conditions for GAT-M.

Table 13.

Differences between NR and WR Among All Conditions for MST-3Stage for GAT-M

Condition	Ability Mean		*t*-test	*SE* Mean		*t*-test
	NRWR	WRNR		NRWR	WRNR	
DPI	0.763	0.992	113.9**	0.977	0.942	3146.5**
AMI	0.472	0.761	111.3**	1.00	0.971	4802.6**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with *df* 9,107

Regarding groups, Figure 9 shows that the estimated abilities were slightly higher in WRNR conditions with DPI in all groups when compared to the groups in NRWR. The same scenario occurred in WRNR conditions with the AMI method; the abilities were larger for all groups in WRNR conditions in relation to NRWR conditions. However, the

size of the group with low ability group shrank in both NRWR and WRNR using the AMI method.

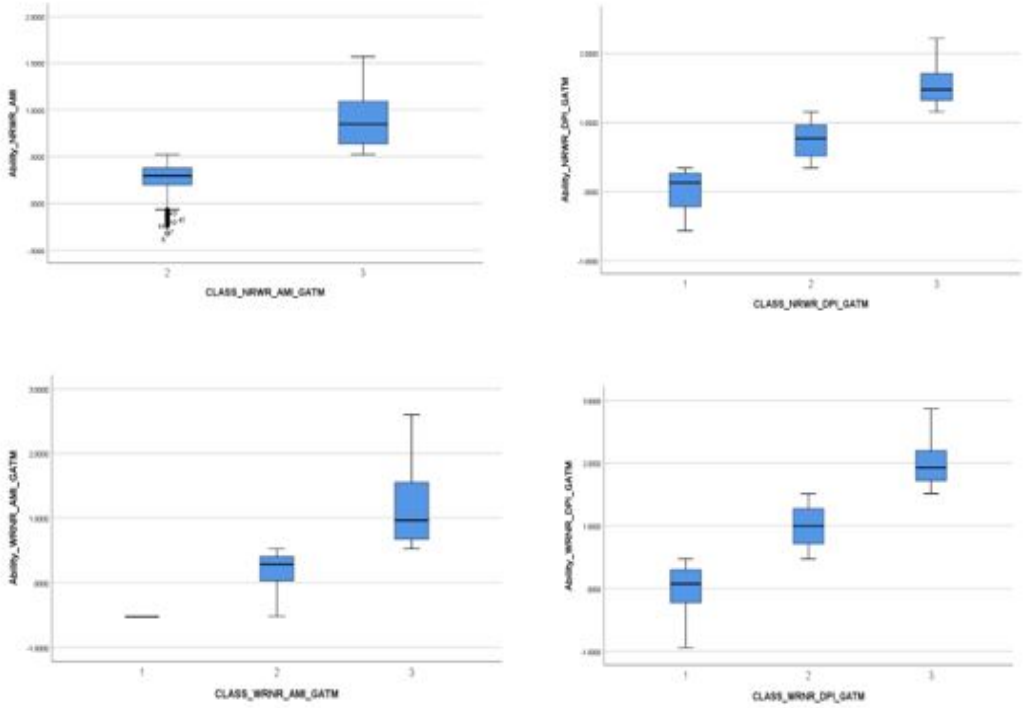

Figure 9. Classes in all conditions in MST-3Stage for GAT-M, NRWR-AMI, and WRNR-AMI (left), NRWR-DPI and WRNR-DPI (right)

GAT-M: Comparison between NR and WR in accuracy of classification decisions. The classification methods that were used in the GAT-V section were also applied to the GAT-M section. The examinees were classified into three classes according to their estimated abilities in MST: high (above average), medium (average), and low (below average) classes for both DPI and AMI conditions.

For MST-2Stage, the percentages of correct classification decisions were generally small in all conditions, ranging from 60.2 to 39% (Table 14). The percentage of correct classifications were higher in NR (60.2%) than in WR (58.2%), using the DPI

method. On the other hand, the WR condition had slightly higher correct classification decisions (40.3%), compared to the NR-AMI condition (39%) with the AMI method. Generally, the incorrect classification decision percentages were high in both NR and WR with the AMI method. Approximately all incorrect decisions were positive false classification decisions (60.6% and 59.5%), wherein examinees were incorrectly classified into classes in MST that were higher than their classes in CLT.

For MST-3Stage, the NRWR and WRNR showed a slight difference in percentages of classification decisions while using the same routing method. The percentage of correct classification decisions was higher in WRNR (63.9%) than in NRWR (62.8%), using the DPI method; conversely, with the AMI method the percentage was higher in NRWR (57.6%) than in WRNR (38.9%). The highest percentage of incorrect classification was a false positive decision with AMI conditions; it was particularly higher in WRNR (61%) than in NRWR (40.2%). It is clear that the cut-scores in AMI conditions impacted the percentage of correct classification in MST by pulling examinees into levels that were higher than their levels in CLT.

Table 14.

Percentage of Correct Classification Decisions in All Conditions of MST GAT-M

Condition	Precision of classification decision	MST-2STAGE NR %	MST-2STAGE WR %	MST-3STAGE NRWR %	MST-3STAGE WRNR %
DPI	CD	60.2	58.2	62.8	63.9
	FP	19.3	19.9	18.2	17.6
	FN	20.5	21.8	19.0	18.5
AMI	CD	39.0	40.3	57.6	38.9
	FP	60.6	59.5	40.2	61.0
	FN	0.4	0.2	2.3	0.2

CD= correct decision, FP= false positive decision, FN= false negative decision

Research Question Two

Does the performance of DPI and AMI routing methods yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V and GAT-M?

The two routing methods, DPI and AMI, were applied to assign examinees into the second stage of MST-2Stage, and into the second and third stages of MST-3Stage for both subtests of the GAT. Comparisons were conducted among all conditions in terms of ability estimates and accuracy in classification decisions. The results are separately presented in two sections, GAT-V and GAT-M.

GAT-V: Comparison between DPI and AMI in estimated ability. A comparison between DPI and AMI conditions in accuracy of estimated ability for GAT-V was conducted separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-V with MST-2Stage. The examinees were assigned to the second stage modules based on their estimated ability in the first routing stage, using DPI and AMI

methods. To assign examinees to the next stage modules, the *DPI* required examinees to be sorted in rank order. The top 30% of the examinees were assigned to the hard module (HM), 40% to the medium module (MM), and the lowest 30% were assigned to the easy module (EM).

The total sample was 9,108 examinees; therefore, the number of cases in the lowest/ highest 30% should have been about 2,732 examinees, who were assigned to the easy/ hard module. The lowest 30% group included examinees who had an estimated ability of -0.478 and lower. Out of the 68 cases that had the same ability level of -0.478, 56 of them must be included in the lowest 30%, while twelve cases must be excluded, according to DPI strategy. However, the decision was made to include all of them in the easy module. The highest 30% group included examinees who had an estimated ability of 0.497 and higher, out of which 107 cases had the same ability level of 0.497; 90 of them must be included in the highest 30%, and 17 must be excluded, according to DPI strategy. The decision was made to include all these cases in the hard module. Therefore, the modules in the second stage were: ***Easy Module,*** including 2,747 examinees (30.2%) with estimated abilities in the routing stage ranging from -1.79 to -0.478; ***Medium Module,*** including 3,612 examinees (39.7%) with estimated abilities in the routing module ranging from -0.475 to 0.495; and ***Hard Module,*** including 2,749 examinees (30.2%) whose estimated abilities in the routing stage ranged from 0.474 to 1.55. See Table 15 for details.

For *AMI,* the examinees were assigned according to their proficiency indicators, which were compared to a performance-based rule, using the test information functions

(TIFs) with IRT. Additionally, the examinees were assigned to three modules based on their estimated proficiency in the routing stage. Next, their scores were compared to two cutoff scores (-0.524 and 0.524), which are indicators of the highest and lowest 30% of the normal continuum distribution of latent ability. *Class 1* included examinees whose estimated ability was -0.524 or lower. *Class 2* included examinees whose estimated ability was between -0.524 and 0.524. *Class 3* included examinees whose estimated ability was 0.524 or higher. Table 15 demonstrates the modules in the second stage. The ***Easy Module*** included 2,522 examinees (27.7%) whose estimated abilities in the routing stage ranged from -1.79 to -0.525; the ***Medium Module*** included 3,977 examinees (43.7%) whose estimated abilities in the routing module ranged from -0.518 to 0.507; and the ***Hard Module*** included 2,609 examinees (28.6%) and their estimated abilities in the routing stage ranged from 0.524 to 1.55.

Table 15.

Descriptive Statistics of the Samples from the Three Modules in the Second Stage in MST-2Stage for GAT-V Using DPI and AMI Methods.

Routing method	Module	# of Cases	%	Estimated Ability on Routing Stage			
				Min	Max	Mean (*SE*)	SD
DPI	Easy	2,747	30.2	-1.79	-0.478	-0.944 (0.007)	0.35
	Medium	3,612	39.7	-0.475	0.495	-0.03 (0.005)	0.27
	Hard	2,749	30.2	0.497	1.55	0.939 (0.006)	0.31
AMI	Easy	2,522	27.7	-1.79	-0.525	-0.984 (0.007)	0.33
	Medium	3,977	43.7	-0.518	0.507	-0.008 (0.005)	0.30
	Hard	2,609	28.6	0.524	1.55	0.963 (0.006)	0.30

The ability means, and the error means were estimated in DPI and AMI for all conditions of the assembling methods NR and WR. The results indicated that there was a slight difference between DPI and AMI conditions regarding the estimated means for ability and standard error in all conditions. The ability means were quite similar in DPI (0.892 and 0.882) and AMI (0.998 and 0.906) in both NR and WR conditions, respectively. However, the difference in ability means was statistically significant even though it was small, $p < 0.001$ with *df* 9,107. The results of *t*-tests, which are presented in Table 16, indicate that the ability means of AMI were significantly higher than the ability means of DPI in both assembling conditions NR and WR.

The standard error (*SE*) means for estimated abilities were also similar in all conditions; however, the *SE* means were slightly smaller with DPI conditions (0.899 and 0.916) than with AMI conditions (0.910 and 0. 924). The *t*-test results show that these differences were statistically significant, $p < 0.001$ with *df* of 9,107, meaning the *SE* associated with the estimated ability was smaller with DPI conditions (see Table 16). According to *t*-test results, the DPI method performed better than the AMI method regarding the accuracy of ability estimation. Figure 10 shows the differences between the *SE* means in DPI and AMI in MST-2Stage for GAT-V. The DPI-NR condition shows the smallest *SE* mean than other conditions of MST-2Stage for GAT-V.

Table 16.

Differences between DPI and AMI Among All Conditions for MST-2Stage of GAT-V

Condition	Ability Mean		*t*-test	*SE* Mean		*t*-test
	DPI	AMI		DPI	AMI	
NR	0.982	0.998	112.1**	0.899	0.910	1254.3**
WR	0.882	0.906	110.2**	0.916	0.924	1343.3**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with *df* 9,107

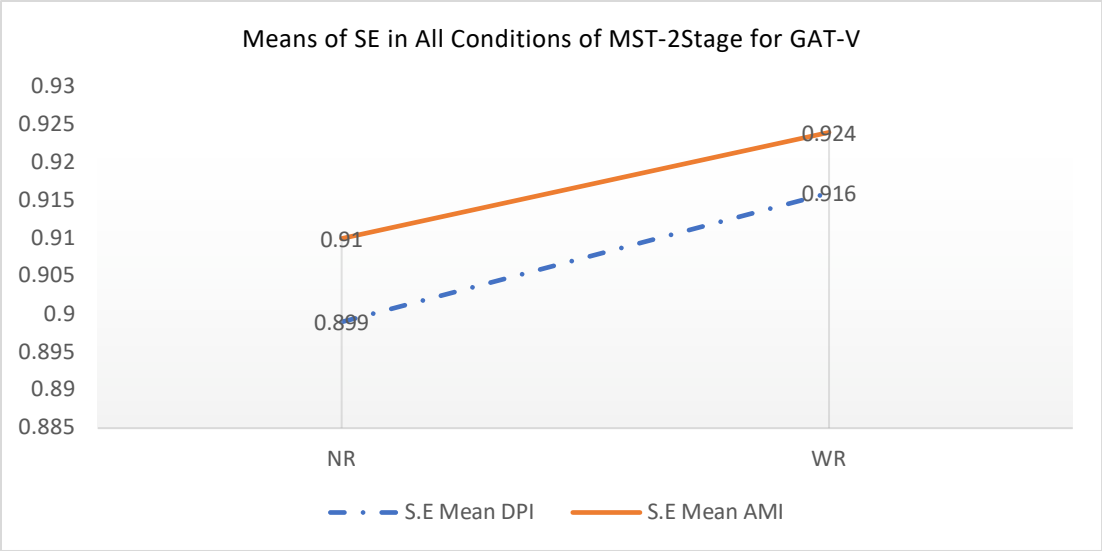

Figure 10. Differences in SE means between DPI and AMI in MST-2Stage for GAT-V

GAT-V with MST-3Stage. Examinees' abilities were estimated in the GAT-V with MST-2Stage for all conditions. First, the examinees were classified using DPI and AMI methods, into three classes (high, medium and low), based on their abilities from stage 2. See Table 17 for details. Next, each examinee was classified into one of nine groups based on his/her class in the first and second stages. The nine groups represented the nine potential paths (see Chapter 2, p. 47). Finally, the examinees' abilities were estimated in the GAT-V with MST-3Stage after their classes were specified.

Table 17.

Descriptive Statistics for the Samples of the Three Modules in the Third Stage in MST-3Stage GAT-V using DPI and AMI Methods

Condition	Modules	# of Cases	%	Estimated Ability on the Second Stage			
				Min	Max	Mean (S.E)	SD
DPI-NRWR	Easy	2,732	30	-1.20	0.677	-0.064 (0.01)	0.54
	Medium	3,636	39.9	0.679	1.458	1.10 (0.003)	0.21
	Hard	2,740	30.1	1.46	2.50	1.87 (0.006)	0.27
DPI-WRNR	Easy	2,741	30.1	-1.486	0.582	-0.036 (0.010)	0.54
	Medium	3,637	39.9	0.583	1.267	0.963 (0.003)	0.19
	Hard	2,730	20	0.301	2.36	1.70 (0.006)	0.30
AMI-NRWR	Easy	588	6.5	-1.15	-0.528	-0.810 (0.007)	0.17
	Medium	1,724	18.9	-0.523	0.523	0.12 (0.007)	0.32
	Hard	6,796	74.6	0.526	2.46	1.38 (0.006)	0.46
AMI-WRNR	Easy	498	5.5	-1.44	-0.529	-0.975 (0.010)	0.23
	Medium	1,893	20.8	-0.521	0.523	0.169 (0.006)	0.24
	Hard	6,717	73.7	0.525	2.34	1.25 (0.005)	0.44

Table 18 provides the estimates of the ability means and standard error means with DPI and AMI for all conditions of assembling in both NRWR and WRNR. There was a difference between the DPI and AMI conditions in terms of estimated means of

ability and standard error in all conditions. The results of MST-3Stage GAT-V show that the estimated ability means were higher with DPI method. The ability means were significantly higher in DPI conditions (1.298 and 1.218) than the ability means in AMI (1.246 and 0.962) for both NRWR and WRNR conditions, respectively. The *t*-test results show that the difference among ability means were statistically significant, $p < 0.001$ with *df* of 9,107.

Similar to the results of *t*-tests in MST-2Stage GAT-V, the *t*-test results in MST-3Stage indicate that the standard error means were smaller with the DPI method (0.862 and 0.882) than the *SE* means with AMI method (0.940 and 0.935), in both conditions NRWR and WRNR. The *t*-test results that are represented in Table 18 shows that the difference among *SE* means was statistically significant, $p < 0.001$ with *df* of 9,107. Accordingly, the performance of the DPI method was slightly better than the performance of the AMI method in terms of the estimation accuracy of ability in the MST, either with two or three stages. Figure 11 shows the differences between DPI and AMI on *SE* means in MST-3Stage for GAT-V. The DPI-NRWR condition shows the smallest *SE* mean compared to other conditions of MST-3Stage for GAT-V.

Table 18.

Differences between DPI and AMI Among All Conditions for MST-3Stage of GAT-V

Assembling Method	Ability Mean		*t*-test	*SE* Mean		*t*-test
	DPI	AMI		DPI	AMI	
NRWR	1.298	1.246	108.5**	0.862	0.940	1258.4**
WRNR	1.218	0.962	123.8**	0.882	0.935	1217.4**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with *df* of 9,107

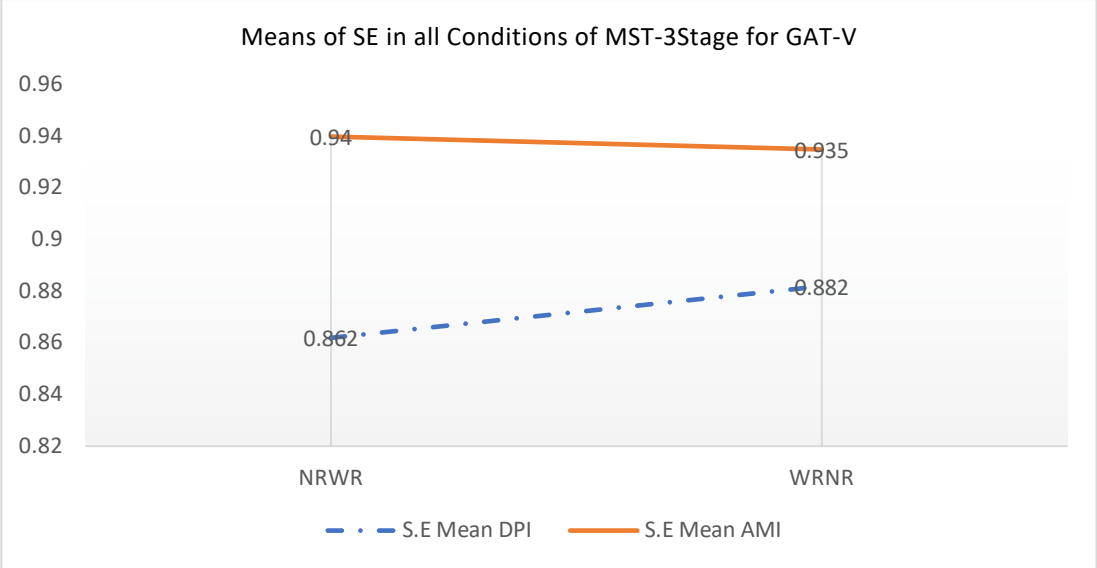

Figure 11. Differences in SE means between DPI and AMI in MST-3Stage for GAT-V

GAT-V: Comparison between DPI and AMI in accuracy of classification decisions. The examinees were classified into three classes/ levels according to their estimated abilities in the MST of GAT-V for both MST-2Stage and MST-3Stage, including the categories of high class (above average), medium class (average), and low class (below average), using DPI and AMI rules.

The percentages of correct classification, presented in Table 9 in research question one above, also include false positive and false negative decisions. The correct classification percentages illustrate cases in which the examinees had the same classes in CLT and MST for GAT-V. False positive decision percentages represent the cases of examinees who had classes in MST that were higher than their classes in CLT for GAT-V, and the percentages of false negative decisions show the percentage of examinees who had lower classes in MST than their classes in CLT.

There were differences between the DPI and AMI conditions in terms of percentage of correct classification decisions, relating to the examinees' classification in CLT of GAT-V. For MST-2Stage, the percentages in DPI (57.9 and 53.5) were higher than in AMI (39.6 and 39.1). Figure 12 shows the percentage of agreement in classification decision of examinees between MST-2Stage and CLT for GAT-V. The DPI-NR had the highest percentage of agreement compared to other conditions.

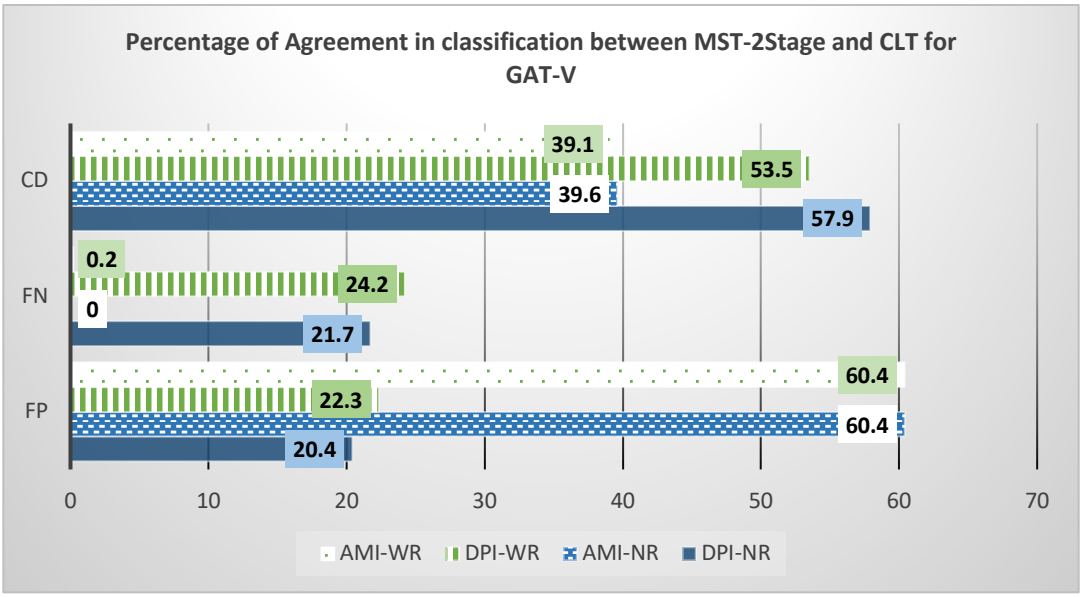

Figure 12. Percentage of agreement in classification decision of examinees between MST-2Stage and CLT for GAT-V

The same scenario occurred in MST-3Stage of GAT-V; the percentages of correct classification decisions were higher in DPI conditions (63.4% and 60.4 %) than in AMI conditions (31.7% and 35.7%) for both NRWR and WRNR conditions, respectively. The use of AMI methods leads to incorrect classification decisions that exceeded 60%; mostly, it was a false positive decision wherein the examinee was classified into classes in MST that were higher than their classes in CLT. Figure 13 shows the percentage of agreement in classification decision of examinees between MST-3Stage and CLT for GAT-V. The DPI-NRWR had the highest percentage of agreement compared to other conditions.

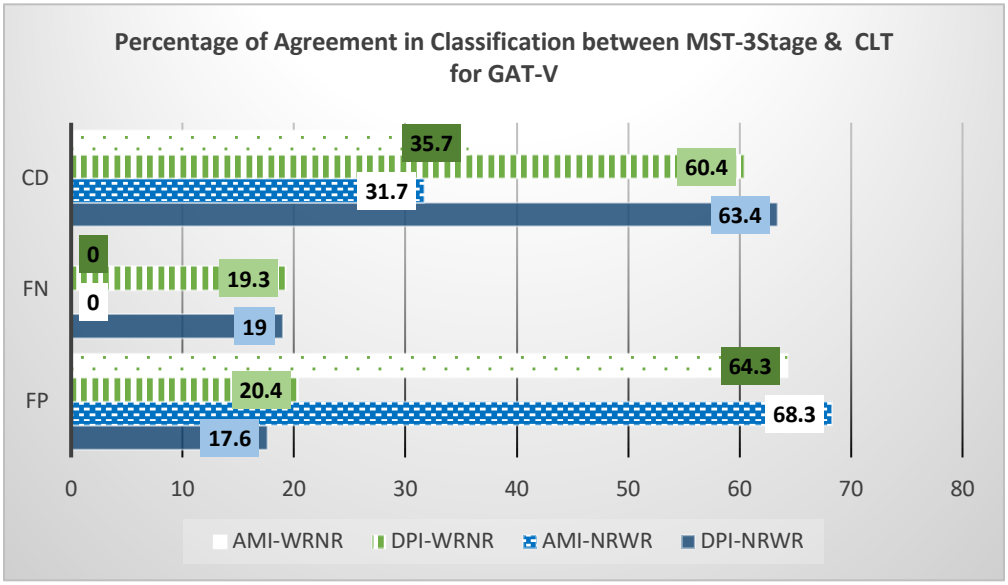

Figure 13. Percentage of agreement in classification decision of examinees between MST-3Stage and CLT for GAT-V

It is interesting to directly compare the DPI classification to the AMI classification in MST, thus, the percentage of agreement between DPI and AMI in MST was computed. The percentage of agreement in classification between the two routing methods was small in all conditions, which was expected because the MST classifications with the DPI method were closer to the CLT classification than the MST classification with AMI. For MST-2Stage, the percentage of agreement between the DPI and AMI was 36.6% in the NR condition while it was 35.6% in the WR condition. For MST-3Stage, the percentage of agreement between the DPI and AMI was 32.7% in the NRWR condition while it was 32.2 % in the WRNR condition.

GAT-M: Comparison between DPI and AMI methods in accuracy of estimated ability. A comparison between DPI and AMI conditions in accuracy of estimated ability for GAT-M was conducted separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-M with MST-2Stage. DPI and AMI methods were used to route examinees to three modules in the second stage, based on their estimated ability in the routing stage. The same percentages and strategies that were used to assign examinees to the next stage in MST for GAT-V were also applied in GAT-M.

For *DPI,* the examinees were sorted in rank order and divided into three classes: top 30% (HM), 40% (MM), and lowest 30% (EM). Therefore, the lowest 30% should include the 2,732 examinees who were assigned to the easy module and the same number of cases must be assigned to the hard module. Furthermore, the lowest 30% group should include examinees who had the estimated ability of -0.568 and lower; there are 143 cases in this ability level, and 88 of them must be placed in the lowest 30%, while 55 cases must be excluded, according to DPI strategy. Therefore, the decision was made to include all of them in the easy module. The highest 30% group should include examinees who had an estimated ability of 0.346 and higher. Since 21 cases had the same ability level, five of them must be included in the highest 30% and assigned to the hard module and 16 must be excluded and assigned to the medium module, according to DPI strategy. The decision was made to include all of them in the medium module.

Table 19 illustrates the classes in the MST-2Stage for GAT-M. There were three modules in the second stage, which are categorized as follows. The ***Easy Module*** included 2,788 examinees (30.6%) whose estimated abilities in the routing stage ranged from -1.43 to -0.57. The ***Medium Module*** included 3,593 examinees (39.4%) whose estimated abilities in the routing module ranged from -0.56 to 0.35. The ***Hard Module***

included 2,727 examinees (29.9%) and their estimated abilities in the routing stage ranged from 0.35 to 1.89.

The examinees were also assigned to three modules according to their scores in the routing stage, which were compared to a performance-based rule, applying *AMI* rules, and using cut-scores of -0.524 and 0.524. The three modules are presented in Table 19. The ***Easy Module*** included 2,906 examinees (31.6%); their estimated abilities in the routing stage ranged from -1.43 to -0.556. The ***Medium Module*** included 3,802 examinees (41.7%); their estimated abilities in the routing module ranged from -0.457 to 0.520. The 2,400 examinees (26.4%) of the ***Hard Module*** had estimated abilities in the routing stage ranging from 0.524 to 1.89.

Table 19.

Descriptive Statistics of the Sample from the Three Modules in the Second Stage of MST-2Stage of GAT-M Using DPI and AMI Methods

Routing method	Modules	# Cases	%	Estimated Ability on the Routing Stage			
				Min	Max	Mean (*SE*)	SD
DPI	Easy	2,788	30.6	-1.434	-0.568	-0.866 (0.005)	0.28
	Medium	3,593	39.4	-0.564	0.346	-0.087 (0.004)	0.24
	Hard	2,727	29.9	0.349	1.892	1.00 (0.009)	0.45
AMI	Easy	2,906	31.9	-1.434	-0.556	-0.853 (0.005)	0.28
	Medium	3,802	41.7	-0.457	0.520	-0.030 (0.004)	0.26
	Hard	2,400	26.4	0.524	1.892	1.08 (0.009)	0.42

The ability means, and the standard error means were estimated in DPI and AMI for all conditions of assembling methods, NR and WR, are presented in Table 20. The results indicate that the differences between the ability means in DPI and AMI conditions were small when using the same assembling methods, either NR or WR. The ability means were comparable in both conditions of DPI and AMI. Regardless of these trivial differences, the ability means in AMI conditions (0.877 and 0.895) were significantly higher than the ability means in DPI conditions (0.852 and 0.870), $p < 0.001$ with *df* 9,107 in both assembling conditions NR and WR, respectively. See Table 20 for *t*-test values.

The difference between the means of standard errors in the DPI and AMI conditions were small, using similar assembling methods, NR and WR. The *SE* means were smaller with AMI conditions (0.947 and 0.935) than with DPI conditions (0.950 and 0.939) in both conditions of NR and WR. The *t*-test results indicated that the differences were statistically significant, $p < 0.001$ with *df* of 9,107. According to *t*-test results, the AMI method performed better than the DPI method regarding the precision of ability estimates. The differences between DPI and AMI on *SE* means in MST-2Stage for GAT-M is displayed in Figure 14. The AMI-WR condition shows the smallest *SE* mean than other conditions of MST-2Stage for GAT-M.

Table 20.

Differences between DPI and AMI Among All Conditions for MST-2Stage of GAT-M

Condition	Ability Mean		*t*-test	*SE* Mean		*t*-test
	DPI	AMI		DPI	AMI	
NR	0.852	0.877	103.9**	0.950	0.947	2100.3**
WR	0.870	0.895	102.9**	0.939	0.935	1773.9**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method

** $p < 0.001$ with df 9,107

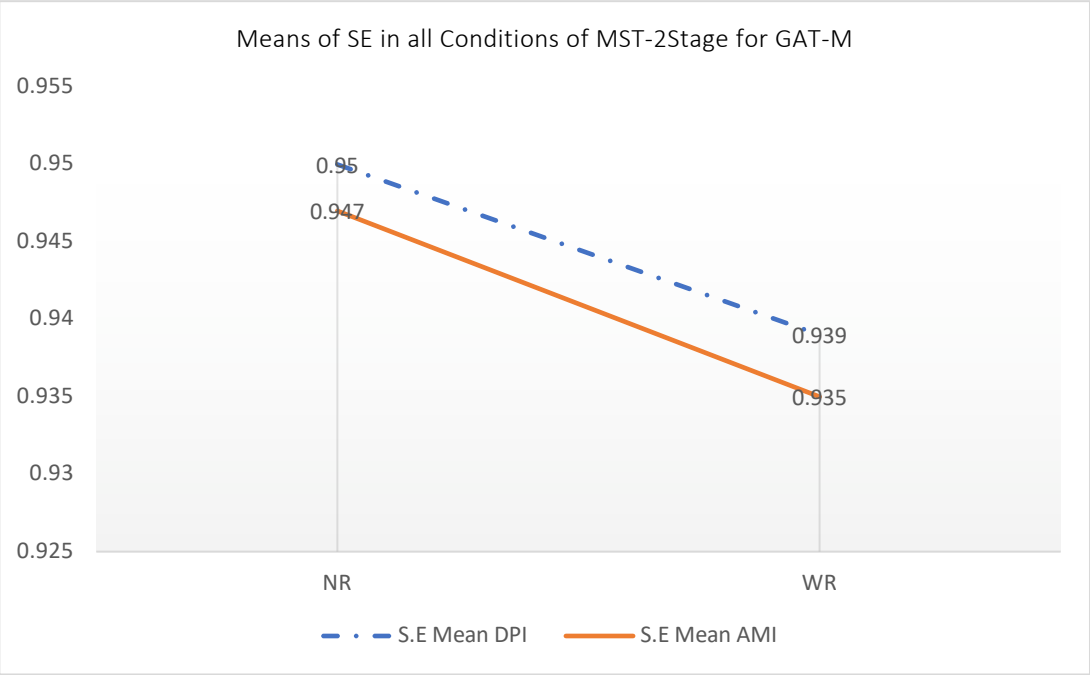

Figure 14. Differences in SE means between DPI and AMI in MST-2Stage for GAT-M

GAT-M with MST-3Stage. The examinees were classified, using DPI and AMI methods, into three classes (high, medium and low), relying on their abilities in stage 2. Table 21 illustrates the classes in all conditions of MST-3Stage GAT-M. The examinees were individually assigned into one of nine groups based on his/her class in the first and second stages. The nine potential paths were explained in Chapter 2, p. 47. As the last

step, the examinees' abilities and the standard errors were estimated in the GAT-M with MST-3Stage.

Table 21.

Descriptive Statistics of the Samples from the Three Modules in the Third Stage in MAT-3Stage GAT-M Using DPI and AMI Methods

Condition	Modules	# Cases	%	Estimated Ability in the Second Stage			
				Min	Max	Mean (*SE*)	SD
DPI-NRWR	Easy	2,733	30	-0.925	0.384	-0.085 (0.007)	0.36
	Medium	3,643	40	0.385	1.25	0.872 (0.004)	0.25
	Hard	2,732	30	1.25	2.61	1.75 (0.007)	0.34
DPI-WRNR	Easy	2,733	30	-1.29	0.589	-0.083 (0.008)	0.42
	Medium	3,700	40.6	0.463	1.26	0.892 (0.004)	0.25
	Hard	2,675	29.4	0.364	2.70	1.81 (0.007)	0.36
AMI-NRWR	Easy	424	4.7	-0.923	-0.524	-0.718 (0.005)	0.11
	Medium	1,731	30	-0.522	0.522	0.076 (0.005)	0.29
	Hard	5,953	65.4	0.525	2.68	1.36 (0.007)	0.52
AMI-WRNR	Easy	585	6.4	-1.31	-0.527	-0.769 (0.010)	0.25
	Medium	2,483	27.3	-0.460	0.523	0.136 (0.005)	0.25
	Hard	6,040	66.3	0.524	2.791	1.37 (0.007)	0.56

The estimates of the ability and standard error means were calculated in DPI and AMI for all conditions of assembling, NRWR and WRNR, and are presented in Table 22. There was a difference between DPI and AMI conditions in terms of estimated means for ability and standard error in all conditions. The mean estimates of ability were higher in the DPI method than in the AMI method. The ability means were significantly higher in DPI conditions (0.763 and 0.992) than the ability means in AMI (0.472 and 0.761), $p <$ 0.001 with *df* 9,107 for both NRWR and WRNR conditions, respectively.

For standard error means, the *t*-test results indicated that the standard error means were significantly smaller with the DPI method (0.977 and 0.942) than the *SE* means with the AMI method (1.00 and 0.971), $p < 0.001$ with *df* 9,107 in both conditions NRWR and WRNR. Therefore, the DPI method performed better than the AMI method in terms of the estimation accuracy of ability in MST for GAT-M with three stages. See Table 22 for details. Figure 15 presents the differences between DPI and AMI on *SE* means in MST-3Stage for GAT-M. The DPI-WRNR condition shows the smallest *SE* mean than other conditions of MST-3Stage for GAT-M.

Table 22.

Differences between DPI and AMI Among All Conditions for MST-3Stage of GAT-M

Assembling Method	Ability Mean		*t*-test	*SE* Mean		*t*-test
	DPI	AMI		DPI	AMI	
NRWR	0.763	0.472	113.9**	0.977	1.00	5643.5**
WRNR	0.992	0.761	115.2**	0.942	0.971	3146.5**

Note. DPI = Defined Population Intervals, AMI =Approximate Maximum Information Method
** $p < 0.001$ with *df* 9,107

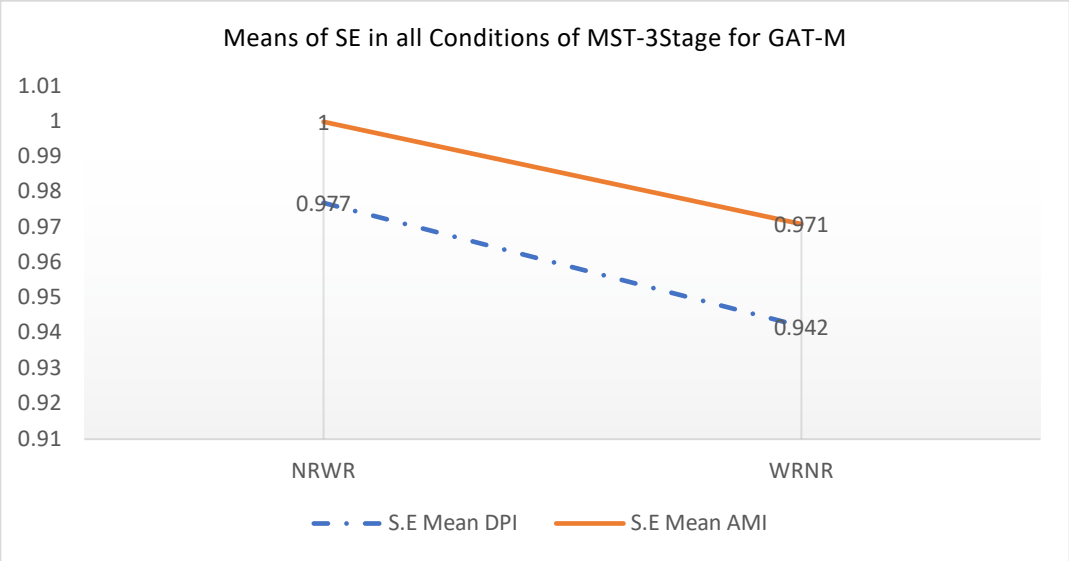

Figure 15. Differences in SE means between DPI and AMI in MST-2Stage for GAT-M

GAT-M: Comparison between DPI and AMI accuracy of classification decisions. The examinees were classified into three levels according to their estimated abilities in GAT-M with MST-2Stage and MST-3Stage: high class (above average), medium class (average), and low class (below average), using DPI and AMI rules.

Table 9 in research question one (above) illustrates the percentages of classification decisions in three statuses: correct, false positive, and false negative classification decisions. The percentage of correct classification represents cases in which the examinees had the same classes in CLT and MST for GAT-M; the false positive decision percentage represents the cases in which the examinees had classes in MST that were higher than their classes in CLT; and the percentage of false negative decisions shows the percentage of examinees who had lower classes in MST than their classes in CLT.

For MST-2Stage, the percentages in DPI (60.2% and 58.2%) were generally higher than in AMI (39% and 40%) for both conditions of assembling, NR and WR. Figure 16 shows the percentage of agreement in classification decision of examinees between MST-3Stage and CLT for GAT-M, and the DPI-NR had the highest percentage of agreement compared to other conditions.

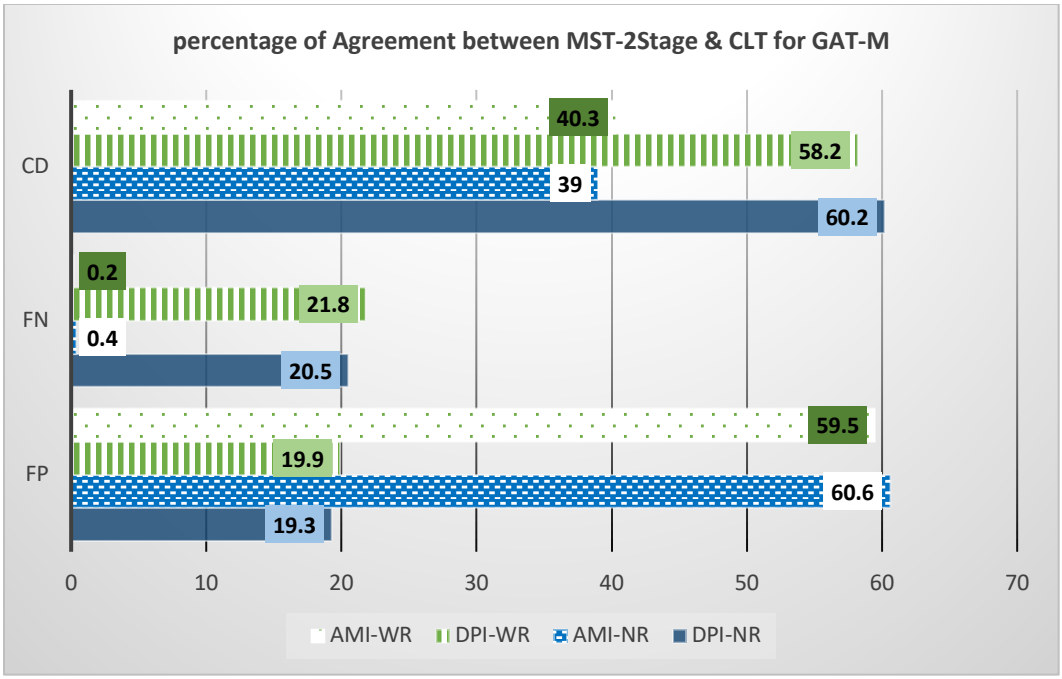

Figure 16. Percentage of agreement in classification decision of examinees between MST-2Stage and CLT for GAT-M

Like MST-2Stage of GAT-M, the MST-3stage shows that the percentage of correct classification decisions was higher in DPI conditions (62.8% and 63.9%) than in AMI conditions (57.6% and 38.9%) for both NRWR and WRNR conditions, respectively. It is notable that the AMI method yielded incorrect classification decisions that exceeded 60%. In most of these cases, the incorrect classification decisions were false positive decisions wherein the examinees were classified into classes in MST higher than their classes in CLT. In conclusion, the AMI method was affected by increasing the

number of stages while the DPI method was independent from the impact of the number of stages in both sections of the GAT. Figure 17 shows the percentage of agreement in classification decision of examinees between MST-3Stage and CLT for GAT-M. The DPI-WRNR had the highest percentage of agreement compared to other conditions.

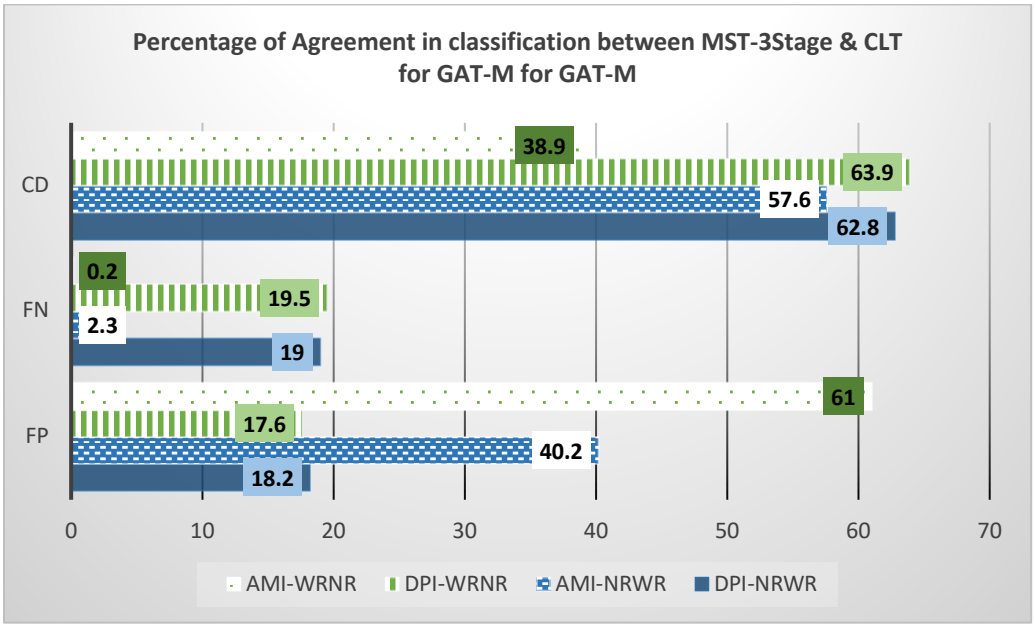

Figure 17. Percentage of agreement in classification decision of examinees between MST-3Stage and CLT for GAT-M

Finally, the percentage of agreement between the DPI and AMI in MST were computed. The percentage of agreement in classification between the two routing methods was small in all conditions. For MST-2Stage, the percentage of agreement between DPI and AMI was 38.8% in the NR condition while it was 39.2% in the WR condition. For MST-3Stage, the percentage of agreement between the DPI and AMI was 44.4% in the NRWR condition while it was 39.4 % in the WRNR condition.

Research Question Three

Does the performance of MST-2Stage and MST-3Stage yield comparable results in terms of ability estimates and accuracy in classification decisions for GAT-V/GAT-M?

Two test designs were developed for MST: MST with two stages and MST with three stages for both sections of the GAT. In all designs, the first stage included nine items, the second stage included three items in each module and the third stage also included three items in each module. Because of a limited item pool, it was impossible to develop MST with NR/WR in the second and third stages; there were too few easy and difficult items. Therefore, the alternative solution was to reverse the assembling method in the third stage. For example, if the assembling method in the second stage was NR, the method in the third stage was WR and in this case the design became NRWR. The designs were separately developed for all conditions of assembling (NR, WR, NRWR, and WRNR), as well as for both conditions of routing methods (DPI and AMI). The aim was to assess the performance of the number of stages in MST, as well as to evaluate the impact of the interaction between routing and assembling conditions in MST-2Stage and MST-3Stage for both sections of GAT.

GAT-V: Comparison between MST-2Stage and MST-3Stage in accuracy of estimated ability. The designs of MST-2Stage and the MST-3Stage for GAT-V are illustrated in Tables 7 and 9. The MST-2Stage included one module in the first stage (medium difficulty), and three modules in the second stage, with different difficulty levels (easy, medium, and hard). In the same manner, the MST-3Stage included one module in the first stage (medium difficulty), and three modules in the second and third

stages with different difficulty levels (easy, medium, and hard). The ability and the standard error means for all conditions were estimated in MST-2Stage and MST-3Stage and presented in Table 23.

The results indicated that the ability means were significantly higher in MST-3Stage (1.30, 1.25, 1.22, and 0.962) than the ability means in MST-2Stage (0.981, 0.998, 0.882, and 0.907), using the same conditions of routing methods and the same assembling methods in the second stage. The *t*-test results show those differences to be statistically significant, $p < 0.001$ and degree of freedom of 9,107. See Table 23 for details.

Table 23.

Differences between MST-2Stage and MST-3Stage in Ability and SE Means Among All Conditions for MST of GAT-V

Condition	Ability Mean		*t*-test	Ability Mean		*t*-test
	NR	NRWR		WR	WRNR	
DPI	0.981	1.30	108.5**	0.882	1.218	123.8**
AMI	0.998	1.25	163.2**	0.907	0.962	141.6**
	SE Mean			*SE* Mean		
	NR	NRWR		WR	WRNR	
DPI	0.897	0.862	1258.4**	0.916	0.881	1343.3**
AMI	0.910	0.940	1217.4**	0.924	0.935	1416.5**

** $p < 0.001$ with *df* 9,107

Regarding the differences between standard error means between MST-2Stage and MST-3Stage, the differences were small. The standard error means for estimated ability were significantly smaller in MST-3Stage (0.862 and 0.881) than the *SE* means in the MST-2Stage (0.897 and 0.916), with the DPI condition, $p < 0.001$ and degree of freedom of 9,107. However, the *t*-test results show that the *SE* means were statistically

smaller in MST-2Stage (0.910 and 0.924) than in MST-3Stage (0.940 and 0.935) using the AMI conditions; the differences were significant, $p < 0.001$ and $df = 9{,}107$. In short, the performance of MST with 2 or 3 stages was enhanced by the type of routing method; the MST-3Stage performed better with the DPI method while the MST-2Stage performed better with the AMI method. T-test values are presented in Table 23.

GAT-V: Comparison between MST-2Stage and MST-3Stage in accuracy of classification decisions. As described in research questions one and two (see Table 9), the examinees were classified into three classes/ levels according to their estimated abilities in MST of GAT-V in both MST-2Stage and MST-3Stage, high, medium, and low class, using DPI and AMI rules.

The percentages of correct classification decisions, when the examinees had the same classes in CLT and MST for GAT-V, were higher in MST-3Stage (63.4% and 60.4%) than MST-2Stage (57.9% and 53.5%) with the DPI method. In contrast, the percentages of correct classification decisions were higher in MST-2Stage (39.6% and 39.1%) than MST-3Stage (31.7% and 35.7%) with the AMI method.

The false positive decision percentages, in which the examinees were classified in classes in MST that were higher than their classes in CLT, were higher in MST-2Stage (20.4% and 22.3%) than in MST-3Stage (17.6% and 20.4), using the DPI method. However, the percentage of false positive decisions was higher in MST-3Stage (68.3% and 64.3%) than in MST-2Stage (60.4% and 60.7%), using the AMI method. As was illustrated in research questions one and two, the use of AMI methods led to incorrect classification decisions, and the highest percentage belonged to the false positive

decisions in which the examinees were classified into classes in MST that were higher than their classes in CLT for GAT-V.

GAT-M: Comparison between MST-2Stage and MST-3Stage in accuracy of estimated ability. The MST-2Stage and the MST-3Stage designs for GAT-M were described above (see Tables 10 and 12 in research question one). The MST-2Stage included one module in the first stage with medium difficulty, and three modules in the second stage with different difficulty levels: easy, medium, and hard. In the same manner, the MST-3Stage included one module in the first stage with medium difficulty, and three modules in the second and third stage with different difficulty levels: easy, medium, and hard. Next, the examinees' abilities were computed for both designs.

Table 24 shows the estimated ability means and the standard error means in MST-2Stage and MST-3Stage for all conditions. The results indicated that ability means were significantly higher in MST-2Stage for NR-DPI (0.851), NR-AMI (0.87), and WR-AMI (0.895) than the ability means in MST-3Stage for NRWR-DPI (0.763), NRWR-AMI (0.472), and WRNR-AMI (0.760), respectively. In the WRNR-DPI condition, the ability mean was higher in WRNR-DPI (0.99) than the ability mean in WR-DPI (0.87). All the *t*-test results show those differences to be statistically significant, $p < 0.001$ and degree of freedom of 9,107.

The differences between standard error means for MST-2Stage and MST-3Stage were statistically significant; $p < 0.001$, 9,107 *df*. The standard error means for estimated ability were significantly smaller in MST-2Stage conditions, NR-DPI (0.950), NR-AMI (0.947), WR-DPI (0.939) and WR-AMI (0.895) than the *SE* means in MST-3Stage

conditions, NRWR-DPI (0.978), NRWR-AMI (1.00), WRNR-DPI (0.942) and WRNR-AMI (0.971), $p < 0.001$. In general, the MST-2Stage performed better than the MST-3Stage in all conditions. T-test values are presented in Table 24.

Table 24.

Differences between MST-2Stage and MST-3Stage in Ability and SE Means Among All Conditions for MST GAT-M

Condition	Ability Mean		*t*-test	Ability Mean		*t*-test
	NR	NRWR		WR	WRNR	
DPI	0.851	0.763	103.9**	0.870	0.992	102.9**
AMI	0.877	0.472	102.5**	0.895	0.760	101.3**
	SE Mean			*SE* Mean		
	NR	NRWR		WR	WRNR	
DPI	0.950	0.978	2138.8**	0.939	0.942	1773.9**
AMI	0.947	01.00	2100.3**	0.935	0.971	1703.9**

** $p < 0.001$ with *df* 9,107

GAT-M: Comparison between MST-2Stage and MST-3Stage in accuracy of classification decisions. The percentages of classification decisions are provided in Table 14 (p.77), wherein the examinees were classified into three classes/levels (high, medium, and low class) according to their estimated abilities in the GAT-M in both MST-2Stage and MST-3Stage, using DPI and AMI rules.

The percentages of correct classification decisions were higher in MST-3Stage for conditions NRWR-DPI (62.8%) compared to NR-DPI (60.2), NRWR-AMI (57.6%) compared to NR-AMI (39.6%), and WRNR-DPI (63.9%) compared to WR-DPI (58.2%). The WR-AMI had slightly high correct classification decision percentage (40.3%) in comparison to the WRNR-AMI condition (38.9%).

The highest percentage of incorrect classification decision was due to a false positive decision with the AMI method, 60.6% and 59.5% in MST-2Stage, while it was 40.2% and 61.0% in MST-3Stage. The percentages of incorrect classification with the DPI method were approximately equally divided between false positive and false negative decisions. For the DPI method, the percentages of false negative decisions, where examinees were incorrectly classified into lower classes in MST than their classes in CLT, were higher in MST-2Stage (20.5% and 21.8%) compared to the false negative percentages in MST-3Stage (19.0% and 18.5%).

Research Question Four

Do the examinees' estimated GAT ability scores on classical linear testing (CLT) significantly relate to the examinees' estimated ability scores on each condition of MST for both GAT-V and GAT-M?

The correlations between the estimated ability of examinees in GAT-CLT and in GAT-MST conditions were conducted for both sections of the GAT. Additionally, the correlations were computed by groups (high, medium, and low) in all conditions for GAT-V and GAT-M. The correlation coefficients were estimated using Pearson coefficients.

GAT-V: Correlations between estimated abilities in CLT and MST. Correlations between MST and CLT regarding the estimated ability of examinees on the GAT-V were calculated separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-V with MST-2Stage. The correlation coefficients between the estimated ability of examinees in CLT and in MST-2Stage conditions were statistically significant for all conditions of MST for the overall correlation, ranging from 0.63 to 0.69 (Table 25). The scatterplots that illustrate the correlations are provided in Figure 18 for all conditions of MST-2Stage. The figures at the left are for DPI conditions, the figures at the right are for AMI conditions, the top figures are for NR conditions, and the bottom figures are for WR conditions.

The correlation coefficients were similar in all conditions; the coefficients were moderately positive, ranging from 0.632 (WR-DPI and WR-AMI) to 0.687 (NR-DPI). The correlation values were different according to group; in all conditions, the low groups had the highest correlation coefficients compared to the others. However, a nonlinear association existed between the estimated ability of examinees in CLT and in MST for the high groups in all conditions. The medium groups had weak positive relations between the estimated ability of examinees in CLT and in MST for all conditions; the coefficients ranged from 0.352 to 3.64.

Table 25.

Correlation Coefficients between Estimated Ability in CLT and in MST-2Stage for GAT-V

Group	NR-DPI	NR-AMI	WR-DPI	WR-AMI
All groups	0.687	0.686	0.632	0.632
Group 1 (Low)	0.615	0.622	0.569	0.577
Group2 (Medium)	0.361	0.364	0.352	0.355
Group 3 (High)	0.091	0.081	0.00	-0.10

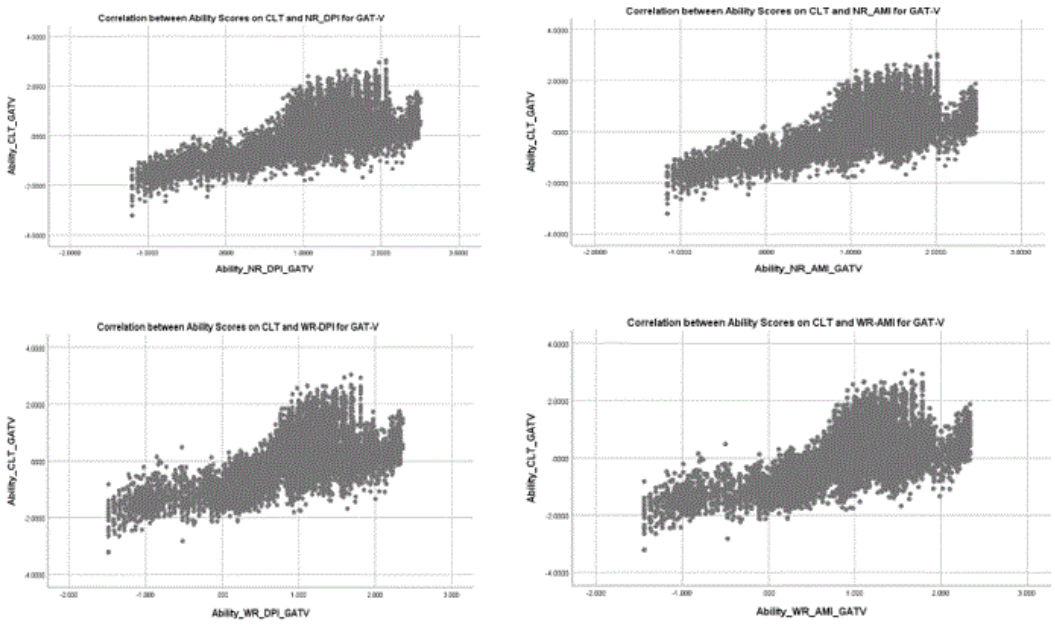

Figure 18. Scatterplots between CLT and MST-2Stage for GAT-V in all conditions, NR-DPI, NR-AMI, WR-DPI, and WR-AMI from the top left to bottom right.

GAT-V with MST-3Stage. The correlations between the estimated ability of examinees in CLT and in MST-3Stage conditions were conducted for all groups individually. The correlations were statistically significant for all conditions of MST-3Stage. The correlation coefficients for all conditions of MST-2Stage of GAT-V are presented in Table 26; the scatterplots are provided in Figure 19 for all conditions of MST-3Stage.

The correlations between the estimated ability of examinees in CLT and in MST-3Stage of GAT-V were strong and positive in all conditions, ranging from 0.75 to 0.87. The correlation coefficients were higher in AMI conditions, NRWR-AMI (0.84) and WRNR-AMI (0.87), than the correlation coefficients in DPI conditions, NRWR-DPI (0.75) and WRNR-DPI (0.76).

The correlation coefficients varied by groups in all conditions. Like the MST-2Stage, the low groups in MST-3Stage had higher correlation coefficients than other groups; the correlations were moderate and positive between ability scores in CLT and in MST-3Stage for low groups, ranging from 0.59 to 0.62. For medium groups in MST-3Stage, the correlation coefficients ranged from weak positive correlation in NRWR-DPI (0.342) to moderate positive in NRWR-AMI (0.56); however, the correlations were higher in WRNR than in NRWR conditions using the same routing methods, either DPI or AMI.

Unlike the high groups in MST-2Stage of GAT-V, the high groups in MST-3Stage AMI conditions had the highest correlation coefficients when compared to the medium groups. The coefficients were moderate and positive in high groups for AMI conditions, NRWR-AMI (0.57) and WRNR-AMI (0.58). The correlation coefficients were weak and positive in high groups for DPI conditions, NRWR-DPI (0.34) and WRNR-DPI (0.25).

Table 26.

Correlation Coefficients Between Estimated Ability in CLT and in MST-3Stage of GAT-V

Group	NRWR-DPI	NRWR-AMI	WRNR-DPI	WRNR-AMI
All groups	0.748	0.841	0.756	0.873
Group 1 (Low)	0.593	0.595	0.623	0.613
Group2 (Medium)	0.342	0.482	0.429	0.557
Group 3 (High)	0.336	0.569	0.252	0.583

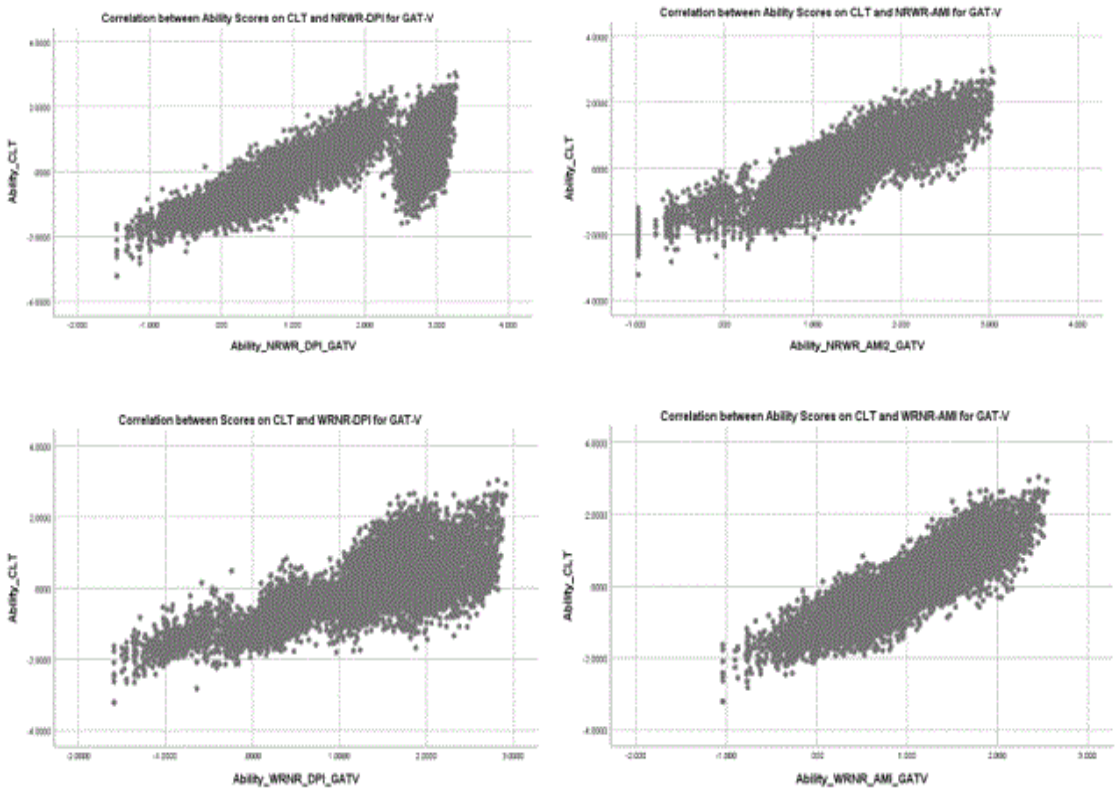

Figure 19. Scatterplots between CLT and MST-3Stage of GAT-V in all conditions, NRWR-DPI, NRWR-AMI, WRNR-DPI, and WRNR-AMI from the top left to bottom right.

GAT-M: Correlations between estimated abilities in CLT and in MST.

Correlations between MST and CLT of the estimated ability of examinees on GAT-M were calculated separately for both panel structure designs of MST: MST-2Stage and MST-3Stage as follows:

GAT-M with MST-2Stage. The correlations between the estimated ability of examinees in CLT and in MST-2Stage conditions were positive and statistically significant for all conditions of MST of GAT-M. Figure 20 includes four plots that represent the correlations between scores in CLT and MST for all conditions of GAT-M. According Table 27, the correlation coefficients were quite similar in all conditions of

MST-2Stage of GAT-M. The correlation coefficients were moderate and positive and ranged from 0.64 (WR-DPI) to 0.67 (NR-DPI).

The correlations were varied among groups in all conditions, and the highest coefficients were moderate and positive (0.47) in low groups for WR-DPI and WR-AMI. The lowest correlation coefficients were 0.26 and 0.32 in high groups for WR-AMI and WR-DPI, respectively. For medium groups, the coefficients were slightly higher in AMI conditions than DPI conditions. Additionally, all correlations between estimated ability in CLT and in MST for all groups were weak and positive, ranging from 0.26 to 0.47.

Table 27.

Correlations Coefficients Between Estimated Ability in CLT and in MST-2Stage of GAT-M

Group	NR-DPI	NR-AMI	WR-DPI	WR-AMI
All groups	0.657	0.673	0.640	0.646
Group 1 (Low)	0.402	0.404	0.465	0.471
Group2 (Medium)	0.381	0.420	0.380	0.409
Group 3 (High)	0.368	0.331	0.323	0.263

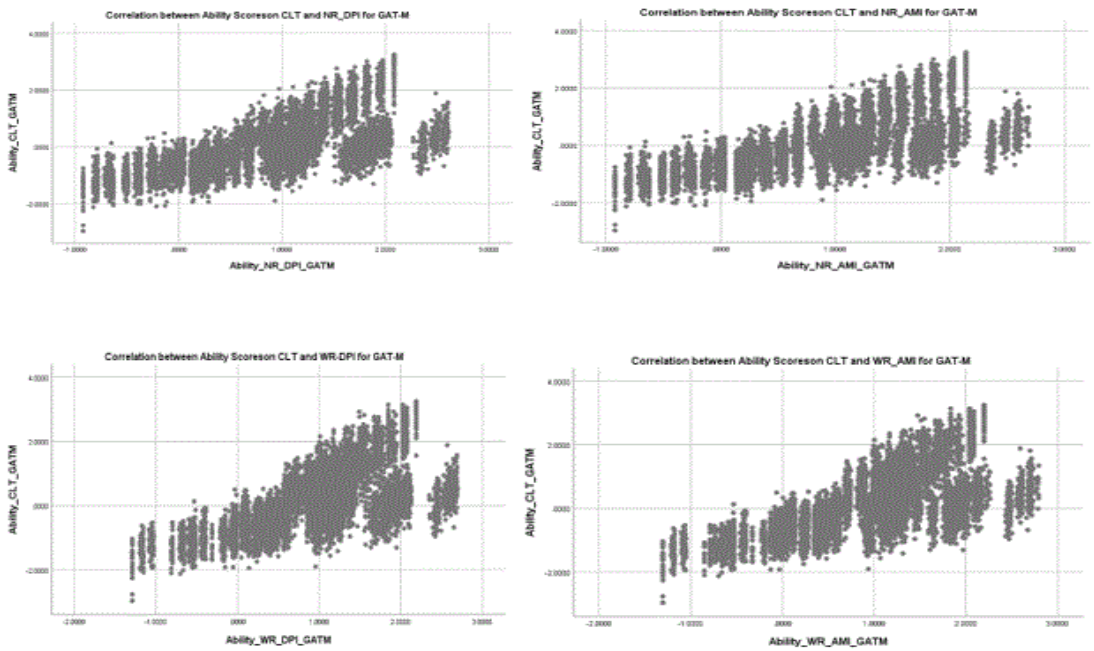

Figure 20. Scatterplots between CLT and MST-2Stage of GAT-M in all conditions, NR-DPI, NR-AMI, WR-DPI, and WR-AMI from the top left to bottom right.

GAT-M with MST-3Stage. The correlations between ability scores in CLT and in MST-3Stage conditions were calculated. The correlations were high and statistically significant for all conditions of MST-3Stage, with values between 0.74 and 0.90. Table 28 shows the correlation coefficients for all conditions of MST-2Stage. The linear positive correlations are obvious in the plots, which are presented as Figure 21.

The correlations between the estimated ability of examinees in CLT and in MST-3Stage were strong and positive in all conditions; the coefficients were higher in AMI conditions, WRNR-AMI (0.90) and NRWR-AMI (0.85), than the correlation coefficients in DPI conditions, WRNR-DPI (0.75) and NRWR-DPI (0.74).

The correlation coefficients were also calculated by group for all conditions of GAT-M and were found to be different among those groups. Unlike MST-2Stage, the

high groups in MST-3Stage had the highest correlation coefficients compared to other groups. Particularly, the correlations between ability scores in CLT and in MST-3Stage for this group were strong and positive (0.80), which is a higher correlation compared to other groups in AMI conditions. The low groups had the lowest correlation coefficients; all were weak and positive and ranged from 0.30 to 0.46. The correlation coefficients in the medium group in MST-3Stage were also weak and positive (0.42 and 0.44 in DPI conditions) and moderate and positive (0.50 and 0.57 in AMI conditions), but higher in the medium groups than in the low groups.

Table 28.

Correlation Coefficients Between Estimated Ability in CLT and in MST-3Stage of GAT-M

Group	NRWR-DPI	NRWR-AMI	WRNR-DPI	WRNR-AMI
All groups	0.741	0.851	0.754	0.899
Group 1 (Low)	0.336	0.305	0.452	0.464
Group2 (Medium)	0.418	0.485	0.442	0.565
Group 3 (High)	0.563	0.802	0.514	0.801

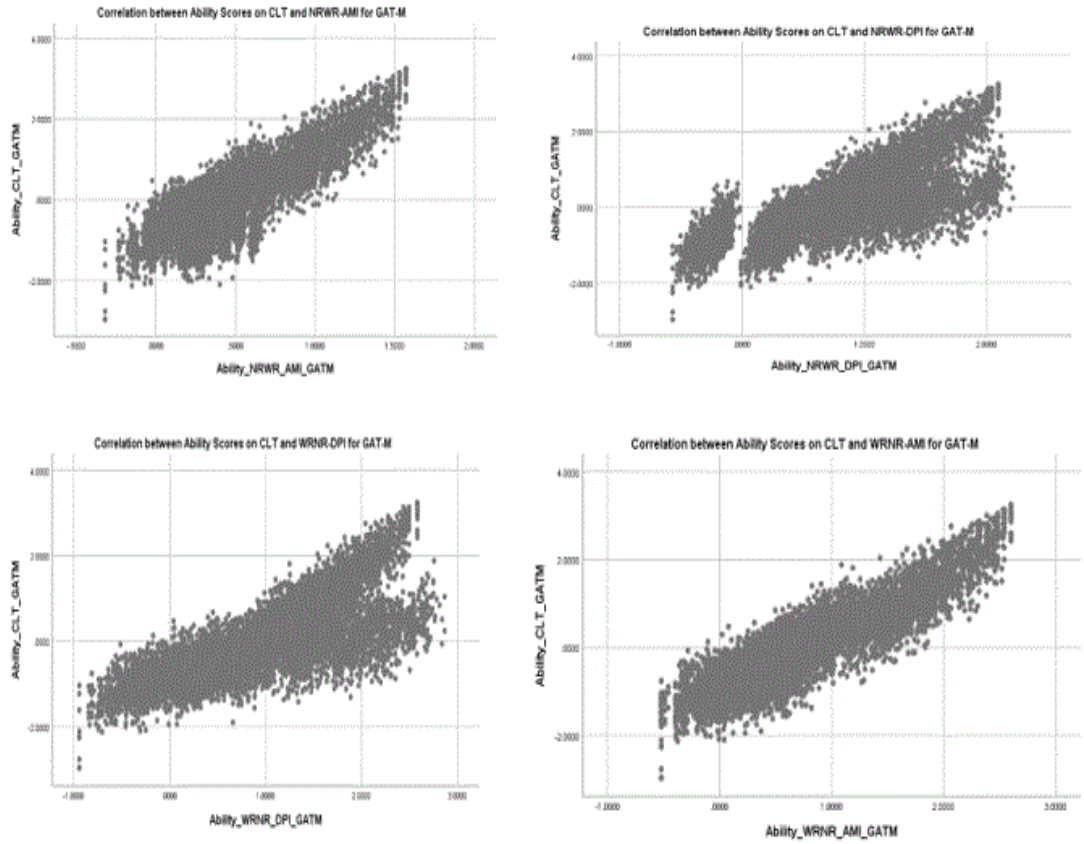

Figure 21. Scatterplots between CLT and MST-3Stage of GAT-M in all conditions, NRWR-DPI, NRWR-AMI, WRNR-DPI, and WRNR-AMI from the top left to bottom right.

Summary

Although the results varied across all conditions for both sections of the GAT, the accuracy of estimated ability was similar in all conditions and the mean differences did not exceed 0.06 *SE*. The results showed noticeable differences regarding the correlations between examinees' scores on CLT and MST conditions. Generally, the MST-3Stage designs showed higher correlations between examinees' scores on CLT and MST than MST-2Stage. Particularly, the AMI conditions of MST-3Stage for both GAT sections had correlations that exceeded $r = 0.70$, while the correlation coefficients in MST-2Stage did not reach 0.70. For GAT-V, the NRWR with the DPI method showed smaller *SE*s and a

correlation of 0.75; however, the WRNR-DPI condition also had a small *SE* and a correlation coefficient of 0.75 for GAT-M.

Regarding the percentage of agreement in classification among conditions, the percentages were low when compared to CLT classification and did not reach 65%. However, the DPI method with MST-3Stage had the highest percentages in all conditions of assembling methods. The AMI conditions had a low percentage of agreement in both sections of the GAT.

The performance of MST-3Stage with the DPI approach was better than the performance of MST-3stage with AMI in term of accuracy of estimated ability (see Figure 22) while MST-3Stage with AMI had high correlations with CLT. NRWR-DPI was better for GAT-V and WR-AMI for GAT-M. The MST-3Stage with AMI condition for both sections of the GAT showed the highest correlations between CLT and MST, but it also showed lowest accuracy of estimated ability and the lowest percentage of classification agreement. In conclusion, the benefit of replacing CLT with MST is high with using MST-3Stage, applying a DPI or AMI approach, and using either a NRWR or a WRNR condition. Figures 23 and 24 illustrate the differences between DPI and AMI regarding the correlation between CLT and MST in GAT-V and GAT-M, respectively. The difference was small between the two methods in MST-2Stage and it became larger in MST-3Stage. The AMI performed better than AMI in MST-3Stage because it showed higher correlations between CLT and MST.

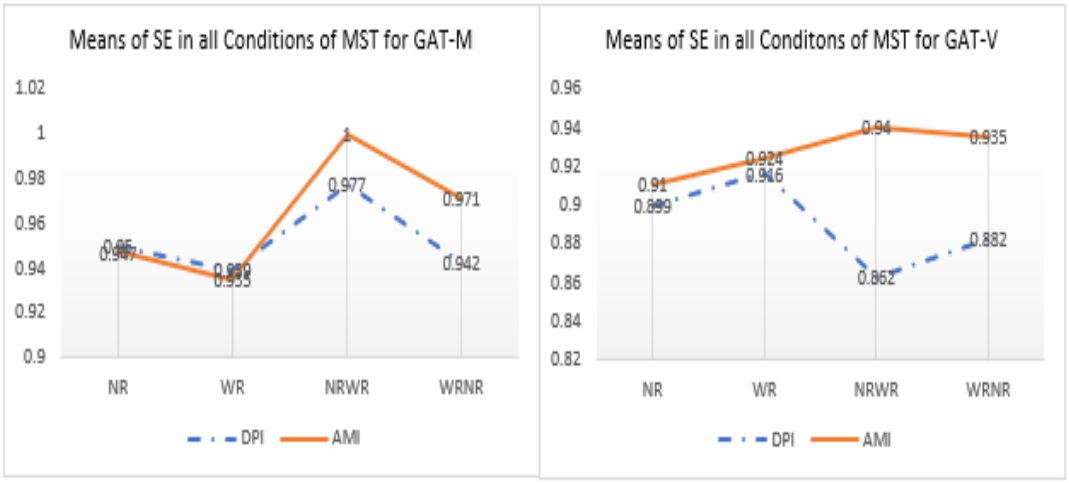

Figure 22. Means of SE for ability estimates in all MST Conditions of GAT-V and GAT-M

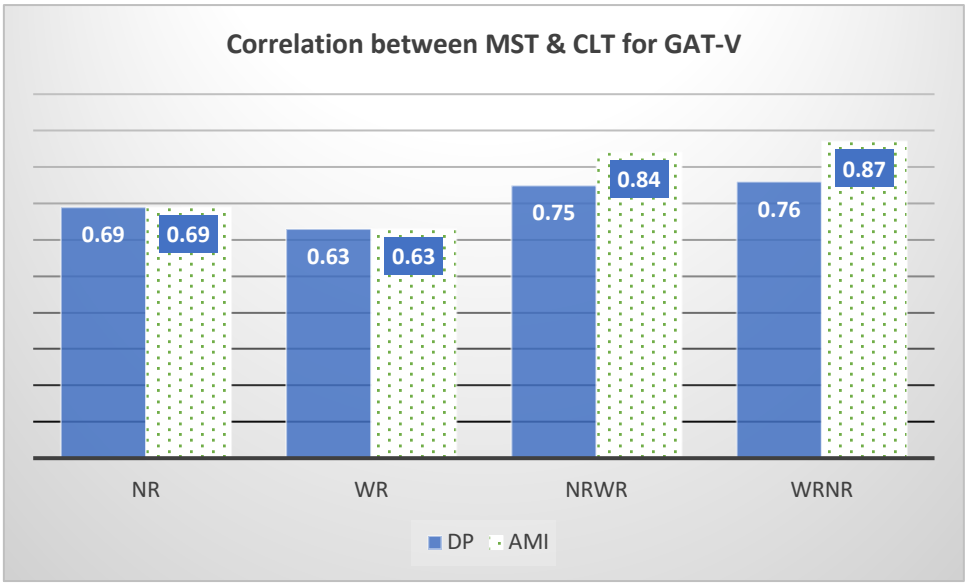

Figure 23. Correlation coefficients on examinees' scores between the MST and CLT for GAT-V

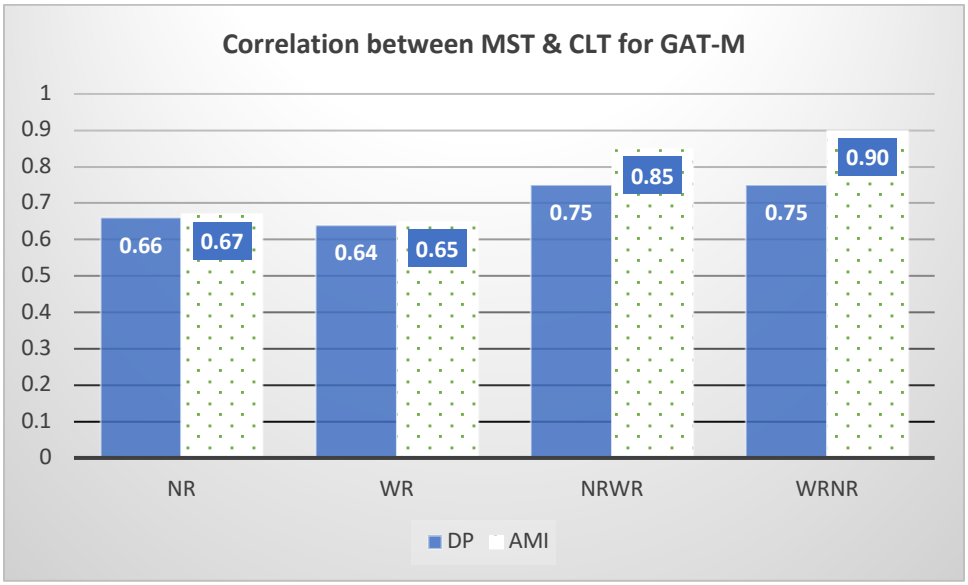

Figure 24. Correlation coefficients on examinees' scores between the MST and CLT for GAT-M

Chapter 4: Discussion

The current study investigated whether multistage testing (MST) could be used as an alternative to classical linear testing (CLT) for the General Aptitude Test (GAT). The use of MST has increased, as its target is to measure a wide range of ability while maintaining control over the content balance of an assessment (Yan, von Davier, & Lewis, 2014). Like CAT, the administration of MST is based on a sequence of adaptive processes; yet, it takes place at the level of each stage, not at the item level, as is the case in CAT (Han, 2013).

The transition from either CLT or CAT to MST has been accomplished for several tests, such as the CPA test (Certified Public Accountants), CTP4 (Comprehensive Testing Program 4), GRE (Graduate Record Examinations), and PI-AAC (Program for International Assessment of Adult Competencies). This transition makes test difficulty more suitable to a student's ability level, in addition to producing easy testing processes for teachers and test coordinators (Yan, von Davier, & Lewis, 2014). The decision to transition from CLT or CAT to MST requires research that addresses numerous questions about MST designs, including appropriate number of modules and stages, module length, module difficulty ranges at each stage, best assembly methods, and routing methods (Yan, von Davier, & Lewis, 2014).

Numerous studies have investigated various MST designs and the effects of different factors on ability estimates and classification decision accuracy. None of these studies combined the effect of the number of stages, assembly method, and routing method using real-world data. Therefore, the current study addressed this gap in the literature and examined the performance of various panel structure designs, routing methods, and assembly methods. The present chapter comprises five subsections: a summary of the study, a summary of the major findings, implications for testing practitioners, limitations of the current study, and recommendations for future research.

Summary of the Study

The primary aim of this study was to compare the performance of two assembly methods (narrow vs. wide range), two routing methods (Defined Population Intervals—DPI— vs. the Approximate Maximum Information method—AMI), and two panel structures (two-stage vs. three-stage) in terms of precision of ability estimates and accuracy of classification for both sections of the GAT. Thus, eight conditions were examined and compared: 2 (assembly condition) * 2 (panel structure) * 2 (routing method) for 2 GAT sections. Four research questions were answered for both the GAT-V and the GAT-M:

a) Does the performance of two assembly conditions (NR, WR) yield comparable results in terms of ability estimates and accuracy in classification decisions?
b) Does the performance of DPI and AMI routing methods yield comparable results in terms of ability estimates and accuracy in classification decisions?

c) Does the performance of two-stage MST and three-stage MST yield comparable results in terms of ability estimates and accuracy in classification decisions?

d) Do the examinees' ability scores on classical linear testing (CLT) correlate with the examinees' scores on each condition of MST?

BILOG-MG3 was used to generate the ability estimates and the standard errors using IRT-2PL. SPSS Version 25 was used for statistical analyses, such as computing descriptive statistics, *t*-tests, correlations, percentages, and classification information.

Major Findings

The main comparisons were individually conducted between two assembly conditions (NR vs. WR), two routing methods (DPI vs. AMI), and two panel structures (MST-2Stage vs. MST-3Stage). The following sections summarize the key results for each research question.

Assembly conditions. The comparison between narrow-range and wide-range conditions yielded trivial differences in the estimated ability means and *SE* means in all conditions. The result of the minor difference between the NR and WR was similar to the result found in Kim and Moses (2014), in which the two assembly methods (small-difference vs. large-difference) had a slight impact.

However, the small difference in *SE* means revealed that the performance of NR was slightly and statistically significantly better than the performance of WR in MST-2Stage for GAT-V. A potential explanation of the statistically significant difference may be due to the large sample size. An alternative explanation is that the examinees faced

certain items that were easer/harder in difficulty level for WR than they faced in the NR condition at the second stage of the MST, which leads to under/overestimates of their abilities in the WR condition. But, the second explanation does not apply to the results of MST-2Stage for GAT-M because the WR condition performed somewhat better than the NR condition in GAT-M. This likely supports the first explanation of the statistical significance in the *t*-test results being a function of a negligible effect by very large sample size.

For MST-3Stage, the limitation of the item pool (there were very few items varying from medium into easy or hard levels of difficulty) forced the assembly methods to be reversed between Stage 2 and Stage 3. Regardless of this limitation, the differences between the assembly methods were minor in terms of the means of estimated ability and *SE*. As stated by Yan, von Davier, and Lewis (2014), the item bank should include sufficient items within the hard and easy levels of difficulty; however, the results of the current study indicate that only very slight differences were found even with an insufficient item bank.

For the GAT-V, the performance of NRWR was better than WRNR with the DPI condition, while the performance of WRNR was better than the NRWR's performance with the AMI method. For GAT-M, the WRNR performed slightly better than NRWR in both routing methods. The different findings regarding effects of the assembly method between the two stages might be because the examinees who were assigned to the WR at stage 2 or stage 3 faced items with higher/lower difficulty levels than they faced in the NR condition, which may lead to inaccurate ability estimates in MST-3Stage with WR,

for either 2 or 3 stages. These multifaceted results in MST-3Stage can be explained by what Wang, Fluegge, & Luecht (2012) and Kim and Moses (2014) found in their studies: complex and simple MST designs performed equally well when the item bank was optimal, which requires including high quality items capable of covering the ability range. In the current study, the item pool was less than ideal.

The classification results support the conclusion of a negligible difference between the NR and WR conditions and show that the NR condition performed slightly better than the WR condition with either DPI or AMI. Like the MST-3Stage results for mean differences, the percentage of agreement in classification was similar in the NR and WR conditions. However, in GAT-V the NRWR excelled with the DPI method, while WRNR excelled with the AMI method; the opposite happened in GAT-M. The results of classification suggest that the differences were not meaningful and could be affected by item bank components.

Routing conditions. The performance of DPI was compared to the performance of AMI in terms of accuracy in ability estimates and in classification decisions. The results indicate that the differences were small between the two routing methods. This conclusion agreed with previous studies that compared routing method performance. Armstrong et al. (2004), as well as Davey and Lee (2011), found similarity between the routing methods in terms of results.

Even though other studies found similarities within routing methods, here the *t*-test results varied between GAT-V and GAT-M. The DPI performed better than AMI in all conditions of MST-2/3Stage for GAT-V and in MST-3Stage for GAT-M. The AMI

method performed better than DPI in MST-2Stage for GAT-M. The differences were small, and did not exceed 0.004 *SE*, indicating that the use of either DPI or AMI had similar impact in terms of accuracy of ability estimates. Results showed that DPI may perform slightly better than AMI, but this conclusion needs additional support as well as examination using an optimal item bank in future studies. The conclusion that can be drawn from this result is that in the AMI method the mean scores were higher than in the DPI condition. Also, the *SE*, in almost all cases, likely differed because of overestimation of ability.

The classification results support this conclusion concerning the AMI method; there were differences between DPI and AMI in terms of classification accuracy. The AMI conditions yielded a high percentage of incorrect classification decisions (exceeding 60% in GAT-V and 40% in GAT-M). Typically, it was a false positive classification for most cases, wherein the examinees were classified into classes in MST that were higher than their classes in CLT.

The results of precision in classification for the current study contradicted the results' found by Kim et al. (2013), which compared different routing methods for four panel structures. They found similarities between routing methods in terms of precision and in classification decisions with the same test length. Item pool limitations likely affected the current results, as previously mentioned; the AMI method pushed examinees into higher classes in MST because the cut-score of the high class was +0.524. and it was easy for most examinees to reach this level of ability since AMI raised their score. The

use of a higher cut-score (such as +1.00) in future studies may change the classification results for the AMI method.

The DPI method showed a higher percentage of agreement between MST and CLT in classification when compared to AMI. The matching in DPI classification between MST and CLT may be enhanced by the large sample size. The large sample helped to achieve a normal distribution of ability scores for CLT (M = 0.0, SD = 0.95). Therefore, the split by rank order worked well for MST, using 30%, 40%, and 30%. It is possible different results would be obtained if a smaller sample size were used.

Panel structure conditions. Comparison between the MST-2Stage and MST-3Stage were conducted for all conditions. The differences between the two panel structures were small but, again, statistically significant.

The *t*-test results show that the performances of MST-2Stage and MST-3Stage were affected by the routing methods. While the MST-3Stage performed better with the DPI method, the MST-2Stage performed better with the AMI method in GAT-V. For GAT-M, the MST-2Stage performed better than the MST-3Stage in all conditions of routing methods. Generally, the agreement percentage between MST and CLT in classification was higher in MST-3Stage than in MST-2Stage. This was expected due to the fact that in MST-2Stage there is only one chance of routing; giving examinees a second opportunity of routing, as in MST-3Stage, can make up for misrouting that may occur in a MST-2Stage.

The lack of items with high and low difficulty levels in the item bank led to a limitation in range of item difficulty for MST-3Stage. However, the results show that a

difference existed between the MST-2Stage and MST-3Stage, which indicated the impact of the number of stages in MST.

In short, the practical results of this study regarding the number of stages in MST suggest that the application of MST-2Stage or MST-3Stage would yield essentially comparable results. However, as mentioned by Yan, von Davier, and Lewis (2014), the construction of MST-2Stage may be simple, but there is a high chance of misrouting since only one routing point exists. This statement was supported by the results of classification in the current study. Thus, MST-3Stage would be recommended for GAT rather than MST-2Stage.

Relationship between examinees' scores on CLT and MST. The results of correlations between the examinees' ability on CLT and MST for both sections of GAT reveal that MST-3Stage had higher correlations with CLT than MST-2Stage with CLT, regarding examinees' scores for all conditions of assembly and routing methods. This conclusion supports the conclusion in the previous section regarding the impact of the number of stages in MST. The correlations were impacted by increasing the number of stages; in other words, increasing the number of indicators yielded an increase in the amount of variability in MST-3Stage compared to MST-2Stage. Thus, the *r* values were greater between MST-3Stage conditions and CLT than the *r* values between MST-2Stage conditions and CLT. As is known. higher variability in scores on X and Y generally leads to greater correlation between X and Y (Goodwin & Leech, 2006).

The correlations between the scores in MST and CLT were also affected by the routing method. The correlations between the scores in CLT and MST were higher in

AMI conditions than in DPI conditions, especially in MST-3Stage. The effect of these routing methods was opposite to the effect of routing methods for classification, wherein the conditions of DPI showed MST in combination with the DPI method had higher percentages in classification agreement with CLT than the case of MST with the AMI method. In the end, the importance for GAT is the correlation more than the classification, since it is the examinees' scores that will be used when they are admitted to college, not their classification.

Correlations by group were weak or moderate at best. This was not unexpected due to the range restriction of scores that impacted the size of *r*. The narrow range of scores shrinks the value of *r* due to shrinking variability (Goodwin & Leech, 2006). Additionally, the size of *r* was affected by the size of the group; for instance, group 3 in MST-3Stage with the AMI method had a large sample size in addition to having the highest correlation between CLT and MST compared to the other groups.

In conclusion, the results of this study support previous research regarding the impact of the assembling and routing methods. This study provided evidence that using either assembly method (NR or WR), either routing method (DPI or AMI), and either two stages or three stages mattered little regarding accuracy in ability estimations. The number of stages mattered when the target was to achieve ability estimates in MST that highly correlated to those estimates from CLT. Giving examinees a second chance of routing to an adaptive point in MST led to estimations of examinees ability that were comparable to CLT.

Implications for Testing Practitioners

Implications can be derived from the results of the current study. Some of these inferences may be new and some of them support the results of prior research; however, all can help testing practitioners. First of all, the implications of the study have the potential to be used differently depending on the content and the purpose of the test. The test may be diagnostic, in which case it is used to classify the examinees, or it may be an ability test aiming to estimate the examinees' scores for certain abilities, regardless of their classification; for either case, additional explanation will be provided later. All the following implications result from the current study, which had limitations that are explained in the next section.

Regarding the methods of assembling and routing, the designs of MST required a wide variety of item difficulty levels to create a test with high accuracy. The study provided evidence that the difference between NR and WR, and between DPI and AMI was of little concern; the difference was so small that the impact was minor.

What *is* a concern involves the test's purpose in addition to the combination of the assembly and routing methods used in conjunction with the different number of stages desired. The combination of NR-AMI in MST-2Stage had the potential to fare well if the test aimed to estimate examinee ability regardless of classification. In contrast, the combination of NR-DPI in MST-2Stage performance would be better for a test that aims to classify examinees.

Another finding of this study that may be of interest to test developers is the number of stages in MST. The MST-3Stage performed better than MST-2Stage regarding

the accuracy of estimates and classification. The combination of the AMI method with either WRNR or NRWR worked better in MST-3Stage, as well as in achieving accurate score estimates. On the other hand, the combination of the DPI method with either NRWR or WRNR worked better in MST-3Stage for classification, which categorizes examinees into distinct levels.

The more satisfactory performance of the DPI method compared to the AMI method in classification is due to the cut-scores used in the AMI method. The cut-scores of 0.524 and -0.524 pulled examinees into levels in MST that were higher than their levels in CLT. This means that the match between CLT and MST in classification appears superior with the DPI method; however, it does not signify that DPI is better than AMI in achieving the best equity and fair classification of examinees. This is an essential aspect to take into consideration when choosing between the application of the two methods, AMI or DPI.

The current study provided evidence that the transition from CLT to MST for GAT is possible; however, a sufficient item bank is necessary. The item bank should have items with an inclusive variety of item difficulty levels, including easy and hard items. The current item bank was designed for the CLT form, which consists of more items of medium difficulty than items with easier or harder levels of difficulty. Further studies are needed to confirm the correct transition from CLT to MST.

Limitations

The main limitation of the current study was the use of a small item pool with an insufficient range from hard to easy in terms of item difficulty. This limitation impacted

the designs of MST for GAT, especially in MST-3Stage. Therefore, this limitation should be taken into account when interpreting the results of the comparison between the NR and WR conditions, and inferences from this study's results must be made with caution. The possibility of the estimated ability being influenced by this limitation was a concern, as was the risk that the analysis may have provided an under/over-estimation of examinee ability, which is another limitation. The content of the GAT was not covered well in MST using the current item pool, some content domains only have items with medium difficulty level (e.g., algebra in GAT-M), other content domains did not have enough hard items (e.g., sentence completion in GAT-V, analysis in GAT-M) because of the lack of breadth in the item pool itself in each domain. Consequently, the groups received test items that did not cover the full range of GAT content, for example in MST for GAT-M, the hard module group in the second stage did not receive any items on algebra because all algebra items in the item pool had medium difficulty. This would impact the accuracy of ability for an examiner, and it may consequently affect the validity of the results. The possibility of differential validity for groups (easy, middle, and hard modules) is one aspect that may be affected by the lack of content coverage. Content validity and other types of validity such as predictive validity of MST for GAT might differ across groups. The groups (modules) that have differential coverage of test content may exhibit different levels of validity which leads to test bias. Therefore, differential validity or statistical adjustments for test bias need to be investigated by future studies that use empirical data.

The sample size was large enough to create separate groups with suitable sizes, but the large sample size may have impacted the results of the analyses conducted

between the targeted conditions. The differences between conditions of MST might not have practical importance though differences were statistically significant, and the results of this study may differ with different sample sizes.

Regarding the analysis, there are few programs available to analyze the MST designs, which was a challenge in the current study. BILOG-MG3 only provides the analysis of MST with two stages, and it has a limited capability to analyze MST with more than two stages. Therefore, the author used the groups analysis in BILOG-MG3 to conduct analyses of the MST-3Stage, which allowed the use of nine groups (representing the nine paths in MST-3Stage), so the calibrations of items could be obtained.

Recommendations for Future Research

The application of MST requires further research into the key factors that may impact MST design features and influence the quality of proficiency estimation as well as classification accuracy. The implications and the limitations of the current study reveal a collection of recommendations that may serve as a guide for future study.

The limitation of the current study's item pool suggests the importance of using an enhanced item bank in future studies when comparing the current combinations among different factors in different conditions of MST. The lack of easy and hard items in the item bank of this study led to interpret the study's results with caution. Additional research is needed to confirm the inferences of this study; future studies are advised to use the current conditions in combination with a sufficient item bank including items of both hard and easy levels of difficulty. It may prove difficult to find an optimal bank with

real-world data; thus, the use of simulation may be an alternative solution even though limitations exist with simulated data.

The AMI method needs supplementary study that uses different cut-scores for routing examinees. Comparisons must be made between the performance results of different cut-off scores. The current study used the 0.524 and -0.524 as cut-scores; other cut-scores, such as +1 and -1, could be used to route examinees to the next stage. Similarly, the use of the DPI method needs additional investigation using different sets of percentages, such as 20%, 60%, and 20%, to assign examinees to the next stage. The use of percentages of 30%, 40%, and 30% worked well in this study for classification accuracy for MST related to CLT. Further information regarding the use of the DPI method with these percentages needs to be investigated with an enhanced item bank and a different sample. In addition, series of pilot studies that aim to determine the most efficient cut-scores of MST for GAT are necessary because it could be good alternative to using other routing methods, and instead of using DPI or AMI.

Because of the conflicting results related to the performance of DPI and AMI regarding estimation and classification accuracy, further studies should be conducted to compare different conditions of routing methods, for both DPI and AMI. Perhaps a pertinent comparison would be among AMI conditions that include two or three possible cut-scores, or a comparison between two or three conditions of the DPI method that include a variety of percentages in a 1-3 or 1-3-3 MST design.

The number of stages and the length of modules in MST are also areas of interest for future studies. The length of the modules in the current study was short: three items in

each module at the second stage in MST-2Stage, as well as at the third stage in MST-3Stage. This led to difficulty in convergence with IRT-3PL because some items were excluded during the calibration process. Therefore, offering a longer module length in the second/ third stages (e.g., 5 to 7 items), as well as an equal length of modules in the routing method, may result in future studies with second and third stages that can be calibrated using IRT-3PL. This may also have an impact on the results of estimation and classification accuracy. Additionally, the comparison between CLT and MST regarding the impact of guessing presents an interesting angle of study, from which future studies may benefit. It is posited that the influence of guessing can be smaller on estimate accuracy in MST than in CLT because the item difficulty levels fit examinee ability levels in MST better than in CLT.

The limited item pool also restricted the number of items in the second stage to only three, which caused a problem for convergence analysis using an IRT-3PL model. Therefore, the IRT-2PL was used in the current study wherein convergence was reached in all conditions of MST. Since the GAT is a multiple-choice test, IRT-3PL may be the more appropriate model and it is recommended that it be investigated in future research.

Finally, the current study used the Bayes Expected *a Posteriori* (EAP) method in test scaling, and then used rescaling so that the posterior latent distribution mean is the location and the SD is the scale. Other scaling method options are available in BILOG-MG3, which can be used and compared in future studies, such as ML (Maximum Likelihood) and Bayes modal Maximum *a Posteriori* (MAP) methods.

References

Alanazi, N. K. J. (2014). *The validity of the General Aptitude Test (GAT) to predict male Saudi students' academic performance in first semester* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3644058)

Alnahdi, G. H. (2015). Aptitude tests and successful college students: The predictive validity of the General Aptitude Test (GAT) in Saudi Arabia. *International Education Studies, 8*(4), 1-6. http://dx.doi.org/10.5539/ies.v8n4p1

Alshumrani, S. A. (2007). *Predictive validity of the General Aptitude Test and high school percentage for Saudi undergraduate students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3266513)

An, X., & Yung, Y.-F. (2014). *Item response theory: What it is and how you can use the IRT procedure to apply it* (Paper SAS364-2014). Cary, NC: SAS Institute, Inc. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28*(3), 147-164. https://doi-org.du.idm.oclc.org/10.1177/0146621604263652

Breithaupt, K., Ariel, A., Hare, D. (2010). Assembling an inventory of multistage adaptiv testing systems. In W. J. Van der Linden and C. A. W. Glas (Eds), *Elements of adaptive testing* (pp.247-266). New York, NY: Springer.

Breithaupt, K., Ariel, A., Hare, D. (2014). The multistage testing approach to the AICPA uniform certified public accounting examinations. In D. Yan, A. A. von Davier, &

C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp.153-168). New York, NY: Chapman and Hall/CRC.

Chen, H., Yamamoto, K., & van Davier, M. (2014). Controlling multistage testing exposure rates in international large-scale assessment. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp.355-369). New York, NY: Chapman and Hall/CRC.

Chuah, S., Drasgow, F., & Luecht, R. (2006). How big is big enough: Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*., *19*(3), 241-255.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE® revised General Test* (ETS Research Report RR-11-26). Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2011.tb02262.x

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications.

Dejong, A., & Molenaar, I. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research, 21*(2), 137-49. https://doi.org/10.1016/0022-3956(87)90014-8

Dimitrov, D. M., & Shamrani, A. R. (2015). Psychometric features of the General Aptitude Test–Verbal Part (GAT-V): A large-scale assessment of high school graduates in Saudi Arabia. *Measurement and Evaluation in Counseling and*

Development, *48*(2), 79-94. https://doi-org.du.idm.oclc.org/10.1177/0748175614563317

Educational Testing Service. (2011). *GRE information and registration bulletin* [Brochure]. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/gre/pdf/gre_info_reg_bulletin.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ.: L. Erlbaum Associates.

Glas, C. (1988). The Rasch model and multistage testing. *Journal of Educational and Behavioral Statistics, 13*(1), 45-52. doi:10.2307/1164950

Goodwin, L. D., Leech, N. L. (2006). Understanding correlation factors that affect size of r. *Journal of Experimental Education, 74*(3), 249-266.

Han, K. T. (2013). *MSTGen*: Simulated data generator for multistage testing. *Applied Psychological Measurement*, *37*(8), 666-668. https://doi-org.du.idm.oclc.org/10.1177/0146621613499639

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383. https://doi-org.du.idm.oclc.org/10.1177/0011000099273004

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–59.

Kean, J., & Reilly, J. (2014). Item response theory. In F. M. Hammond, J. F. Malec, T. G. Nick, & R. M. Buschbacher (Eds.), *Handbook for clinical research: Design,*

statistics and implementation (pp. 195-198). New York, NY: Demos Medical Publishing.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3315089)

Kim, H. Y., & Kim, H. J. (2018, April). *Multistage testing routing designs: Adjacent vs. nonadjacent*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Kim, J., Chung, H., Park, R., & Dodd, B. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods, 45*(4), 1087-1098. https://doi.org/10.3758/s13428-013-0316-3

Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (ETS Research Report RR-14–01). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12000

Kim, S., Moses, T., & Yoo, H. (2015). A Comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement JEM., 52*(1), 70-79.

Kline, R. B. (2010). *Principles and Practices of Structural Equation Modeling*, (3rd Ed.). New York, NY: The Guilford Press.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*(3), 557-575. doi:10.1007/S11336-013-9317-5

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: L. Erlbaum Associates.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249. https://doi-org.du.idm.oclc.org/10.1111/j.1745-3984.1998.tb00537.x

National Center for Assessment in Higher Education. (2017). *General aptitude test (GAT).* Retrieved from http://beta.qiyas.sa/en/Exams/Education/GeneralAbilities/Pages/default.aspx

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Samejima, F. (1997). Graded response model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Sideridis, G. D., & Tsaousis, I., (2013). *DIF analysis for item and test on the NCA tests the General Ability Test* (TR035). The National Center for Assessment in Higher Education. Retrieved from http://beta.qiyas.sa/ar/Researchers/Research-Studies/Pages/reserchingtopics.aspx

Wang, C., Zheng, Y., & Chang, H. H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika*, *79*(1), 154-174. doi: 10.1007/s11336-013-9356-y

Wang, X., Fluegge, L., & Luecht, R. (2012, April). *A large-scale comparative study of the accuracy and efficiency of ca-MST*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*(1), 41-58. https://doi.org/10.1177/0013164406288164

Weissman, A. (2014). IRT-Based multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp.153-168). New York, NY: Chapman and Hall/CRC.

Wentzel, C., Mills, C. M., & Meara, K.C. (2014). Transitioning a K-12 assessment from linear to multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp.355-369). New York, NY: Chapman and Hall/CRC.

Yan, D., von Davier, A. A, & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. New York, NY: Chapman and Hall/CRC.

Zenisky, A. L., & Hambleton, R. K. (2004, April). *Effects of selected multi-stage test design alternatives on credentialing examination outcomes.* Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Zhang, J. (2013). A Procedure for dimensionality analyses of response data from various Test Designs. *Psychometrika,78*(1), 37-58. https://doi.org/10.1007/s11336-012-9287-z

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104-118. https://doi.org/10.1177/0146621614544519

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012, September). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes* (ACT Research Report 6). Retrieved from https://pdfs.semanticscholar.org/049b/54fca400fb73116c438a1f834adcc349f07e.pdf

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika, 80*(1), 65-84. doi: 10.1007/s11336-013-9369-6

Appendix A

Item Parameter Estimates for GAT

Table A1.

Item parameters of the Item Pool for GAT-V

# ITEM	ITEM ID	a[1]	d[2]	LEVEL	# ITEM	ITEM ID	a[1]	d[2]	LEVEL
ITEM01	VAN	0.528	-1.057	M	ITEM01	VCA_H	0.419	1.385	M
ITEM02	VAN_A	0.931	-0.498	M	ITEM02	VCA_I	0.541	0.188	M
ITEM03	VAN_B	0.406	-0.548	M	ITEM03	VSC	0.743	-0.831	M
ITEM04	VAN_C	0.485	-0.165	M	ITEM04	VSC_A	0.167	2.131	H
ITEM05	VAN_D	0.432	-0.398	M	ITEM05	VSC_B	0.46	-0.594	M
ITEM06	VAN_E	0.367	1.654	M	ITEM06	VSC_C	0.385	0.513	M
ITEM07	VAN_F	0.12	3.944	H	ITEM07	VRC_A	0.524	-0.751	M
ITEM08	VAN_G	0.396	0.966	M	ITEM08	VRC_B	0.228	-2.592	E
ITEM09	VAN_H	0.979	-1.124	E	ITEM09	VRC_C	0.201	2.436	H
ITEM10	VAN_I	0.468	-1.513	E	ITEM10	VRC_D	0.398	-0.76	M
ITEM11	VAN_J	0.658	-0.313	M	ITEM11	VRC_E	0.597	-1.032	E
ITEM12	VAN_K	0.506	0.413	M	ITEM12	VRC_F	0.543	-0.252	M
ITEM13	VAN_L	0.147	1.113	M	ITEM13	VRC_G	0.453	1.041	M
ITEM14	VAN_M	0.504	0.894	M	ITEM14	VRC_H	0.118	4.052	H

ITEM15	VAN_N	0.49	0.752	M	ITEM15	VRC_I	0.079	-2.048	E
ITEM16	VAN_O	0.454	0.512	M	ITEM16	VRC_J	0.229	1.146	M
ITEM17	VCA	0.788	-1.171	E	ITEM17	VRC_K	0.457	-0.047	M
ITEM18	VCA_A	0.559	-0.819	M	ITEM18	VRC_L	0.19	4.664	H
ITEM19	VCA_B	0.471	-0.761	M	ITEM19	VRC_M	0.699	-0.71	M
ITEM20	VCA_C	0.52	-0.195	M	ITEM20	VRC_N	0.818	-2.043	E
ITEM21	VCA_D	0.443	0.359	M	ITEM21	VRC_O	0.344	0.045	M
ITEM22	VCA_E	0.382	-0.796	M	ITEM22	VRC_P	0.555	-0.108	M
ITEM23	VCA_F	0.404	0.277	M	ITEM23	VRC_Q	0.245	2.158	H
ITEM24	VCA_G	0.378	0.781	M	ITEM24	VRC_R	0.414	0.849	M

a^1 = item discrimination, d^2 = item difficulty, E= easy, M= medium, H= hard.

Table A2.

Item parameters of the Item Pool for GAT-M

# ITEM	ITEM ID	a^1	d^2	LEVEL	# ITEM	ITEM ID	a^1	d^2	LEVEL
ITEM01	MAR	0.806	-1.879	E	ITEM23	MGE_G	0.447	0.495	M
ITEM02	MAR_A	0.382	1.631	H	ITEM24	MAN	0.33	1.698	H
ITEM03	MAR_B	0.593	-0.234	M	ITEM25	MAN_A	0.411	-0.006	M
ITEM04	MAR_C	0.451	0.429	M	ITEM26	MAN_B	0.51	0.321	M
ITEM05	MAR_D	0.431	0.738	M	ITEM27	MAN_C	0.598	-0.987	E

ITEM06	MAR_E	0.254	0.645	M	ITEM28	MAN_D	0.685	-1.379	E
ITEM07	MAR_F	0.522	0.85	M	ITEM29	MAN_E	0.496	-0.384	M
ITEM08	MAR_G	0.24	2.91	H	ITEM30	MAN_F	0.407	-1.09	E
ITEM09	MAR_H	0.826	-0.029	M	ITEM31	MAN_G	0.425	0.288	M
ITEM10	MAR_I	0.686	-0.294	M	ITEM32	MAL	0.413	0.874	M
ITEM11	MAR_J	0.688	-0.286	M	ITEM33	MAL_A	0.613	-0.254	M
ITEM12	MAR_K	0.31	-0.477	M	ITEM34	MAL_B	0.531	0.267	M
ITEM13	MAR_M	0.486	0.457	M	ITEM35	MAL_C	0.498	0.257	M
ITEM14	MAR_N	0.431	1.015	M	ITEM36	MCO	0.181	-1.149	E
ITEM15	MAR_O	0.465	1.537	H	ITEM37	MCO_A	0.154	0.13	M
ITEM16	MGE	0.472	-0.844	E	ITEM38	MCO_B	0.783	0.537	M
ITEM17	MGE_A	0.421	-0.905	E	ITEM39	MCO_C	0.267	0.846	M
ITEM18	MGE_B	0.558	0.594	M	ITEM40	MCO_D	0.418	0.089	M
ITEM19	MGE_C	0.356	0.515	M	ITEM41	MCO_E	0.22	0.931	M
ITEM20	MGE_D	0.45	1.144	M	ITEM42	MCO_F	0.567	0.719	M
ITEM21	MGE_E	0.416	1.254	M-H	ITEM43	MCO_G	0.26	2.274	H
ITEM22	MGE_F	0.518	0.67	M					

a^1 = item discrimination, d^2 = item difficulty, E= easy, M= medium, H= hard.

Appendix B

Test Information Curves

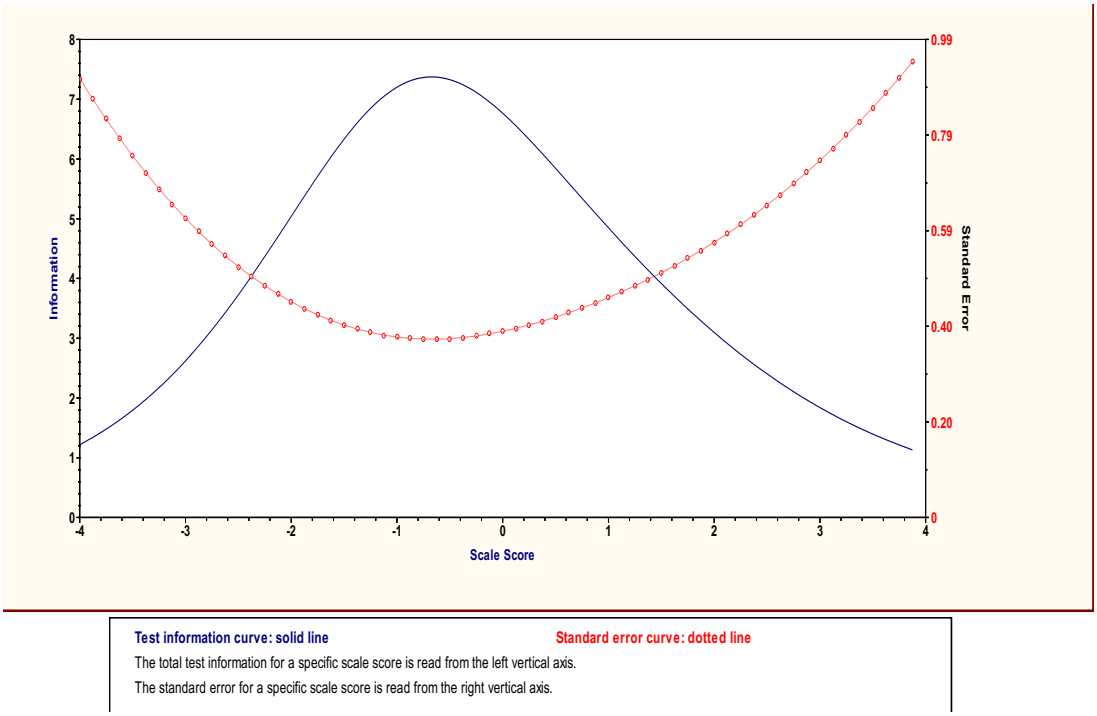

Figure B1. Test Information Curve of CLT for GAT-V Using IRT with 2PL

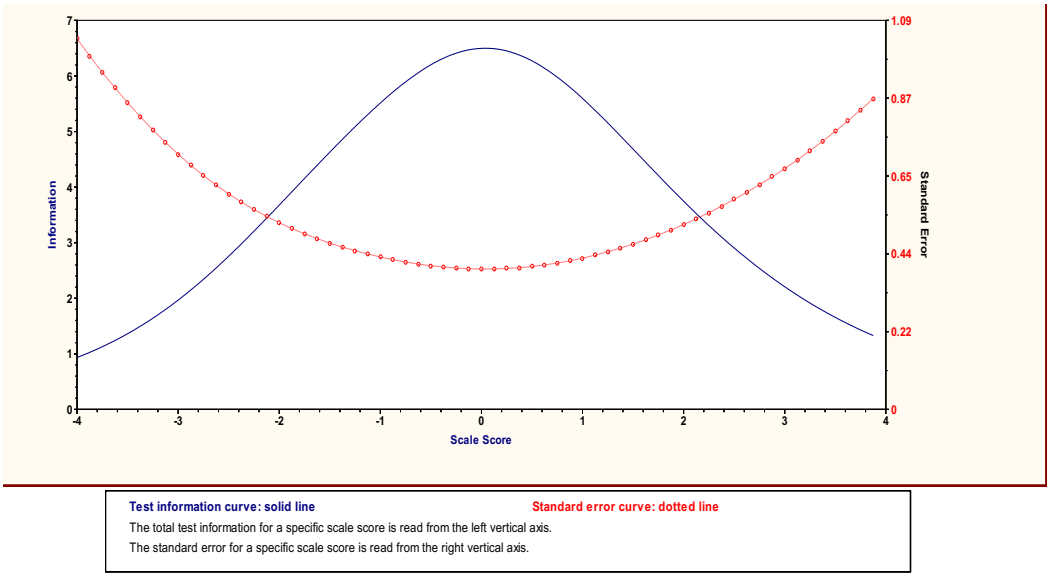

Figure B2. Test Information Curve of CLT for GAT-M Using IRT with 2PL

Appendix C

Evidence for unidimensionality of GAT-V

Table C1. Summary of Model Fit for Single Factor Model for GAT-V

Statistic	Result	Thresholds*
X^2_m	4995.88	Insignificant at a 0.05
df_m	1080	
P	<0.001	
RMSEA (90%CI)	0.0.020 (0.019 - 0.021)	=<0.05(<0.05, <0.10)
SRMR	0.0204	=<0.08
GFI	0.97	=>0.90
AGFI	0.97	=>0.90

The single factor model was programmed using AMOS with 48 items. Table 1 shows the fit indices for single factor model. The chi-square was significant and that was expected with the large sample because this statistical significance test is sensitive to sample size.

The results for the Root Mean Square Error of Approximation RMSEA, which is scaled as a badness-of-fit index, was 0.020 which is smaller than threshold 0.05 and the lower bound of the 90% confidence interval for this statistic was 0.019 < 0.05 and upper bound was 0.021 < 0.10 (Kline, 2010), thus, the adequate fit hypothesis was not rejected, $P > 0.05$.

The value of the Standardized Root Mean Square Residual (SRMR) that measures the mean absolute correlation residual and finds the overall difference between observed and predicted correlations, was 0.0204, smaller than Hu and Bentler's recommended threshold of 0.08 (Kline, 2010). This implies the model had reasonable fit. The values of the Goodness-of-fit (GFI) statistic was 0.97 and the Adjusted Goodness-of-fit (AGFI) statistic was also 0.97; these values are greater than recommended threshold of 0.90,

indicating a well-fitting model (Hopper et al., 2008; Kline, 2010). Overall, the results of fit statistics indicate that the data fit a single factor model well and the unidimensionality of GAT-V was supported.

Appendix D

Evidence for unidimensionality of GAT-M

Table D1. Summary of Model Fit for Single Factor Model for GAT-M

Statistic	Result	Thresholds*
X^2_m	3447.8	Insignificant at a 0.05
df_m	860	
P	<0.001	
RMSEA (90%CI)	0.0.018 (0.018 - 0.019)	=<0.05(<0.05, <0.10)
SRMR	0.019	=<0.08
GFI	0.98	=>0.90
AGFI	0.98	=>0.90

The single factor model was programmed using AMOS with 48 items. Table 1 shows the fit indices for single factor model. The chi-square was significant and that was expected with the large sample because this statistical significance test is sensitive to sample size.

The results for the Root Mean Square Error of Approximation RMSEA, which is scaled as a badness-of-fit index, was 0.018 which is smaller than threshold 0.05 and the lower bound of the 90% confidence interval for this statistic was 0.018 < 0.05 and upper bound was 0.019 < 0.10 (Kline, 2010), thus, the adequate fit hypothesis was not rejected, $P > 0.05$.

The value of Standardized Root Mean Square Residual (SRMR) that measures the mean absolute correlation residual and finds the overall difference between observed and predicted correlations, was 0.019, smaller than Hu and Bentler's recommended threshold of 0.08 (Kline, 2010). This implies the model had reasonable fit. The values of the Goodness-of-fit (GFI) statistic was 0.98 and the Adjusted Goodness-of-fit (AGFI) statistic was also 0.98; these values are greater than recommended threshold of 0.90,

indicating a well-fitting model (Hopper et al., 2008; Kline, 2010). Overall, the results of fit statistics indicate that the data fit a single factor model well and the undimensionality of GAT-M was supported.