Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 60

A Comprehensive Review of Research on Multistage Testing

Hyung Jin Kim, Won-Chan Lee,
Shumin Jing[*], *Huan Liu*[*],
Rabia Karatoprak Ersen[*], *Kuo-Feng Chang*[*], *Min Liang*[*]

December 2025

Hyung Jin Kim is Research Scientist, National Council of State Boards of Nursing (email: hyungjin-kim@uiowa.edu). Won-Chan Lee is Director, CASMA, College of Education, University of Iowa (email: won-chan-lee@uiowa.edu). *Shumin Jing, Huan Liu, Rabia Karatoprak Ersen, Kuo-Feng Chang, and Min Liang were Research Assistants, CASMA, College of Education, University of Iowa.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Web: https://education.uiowa.edu/centers/casma

Contents

LIST OF ABBREVIATIONS

ACT American College Testing
AICPA American Institute of Certified Public Accountants
AIG Automatic Item Generation
AMI Approximate Maximum Information
ATA Automatic Test Assembly
ca-MST Computer-Adaptive Multistage Testing
CAST Computerized Adaptive Sequential Testing
CAT Computerized Adaptive Testing
CBT Computer-Based Testing
CD-MST Cognitive Diagnostic Multistage Testing
CLT Classical Linear Testing
CML Conditional Maximum Likelihood
CPA Certified Public Accountants
CR Constructed Response
CSEM Conditional Standard Error of Measurement
CTT Classical Test Theory
DIF Differential Item Functioning
DPI Defined Population Intervals
EAP Expected a Posteriori
FPC Fixed Parameter Calibration
G-DINA Generalized Deterministic Input, Noisy, and Gate
GPCM Generalized Partial Credit Model
GRE Graduate Record Examinations
ICC Item Characteristic Curves
IRT Item Response Theory
LFT Linear Fixed Length Test
LPFT Linear Parallel-Form Test
LSAT Law School Admission Test
MAD Mean Absolute Difference
MAE Mean Absolute Error
MAP Mode a Posteriori
MAR Missing At Random
MC Multiple-Choice
MCAR Missing Completely At Random
MCAT Medical College Admission Test
MFI Maximum Fisher's Information
MFS Multiple-Form Structure
MIP Mixed Integer Programming
MIRT Multidimensional Item Response Theory
MLE Maximum Likelihood Estimation
MML Marginal Maximum Likelihood
MNAR Missing Not At Random
MST Multistage Test/Testing
NAEP National Assessment of Educational Progress
NC Number-Correct
NWADH Normalized Weighted Absolute Deviation Heuristic
OMST On-the-Fly Multistage Adaptive Testing

P&P Paper-and-Pencil
PCM Partial Credit Model
PIAAC Programme for the International Assessment of Adult Competencies
PL Parameter Logistic
RMSE Root Mean Square Error
SAT Suit of Assessments
SD Standard Deviation
SE Standard Error
SEE Standard Error of Estimates
SEM Standard Error of Measurement
TCC Test Characteristic Curves
TIF Test Information Functions
TRT Testlet Response Theory
UIRT Unidimensional Item Response Theory

Acknowledgements

This literature review was conducted with support from the College Board through a contractual agreement. We gratefully acknowledge their financial support, which made the preparation of this review possible. The authors are solely responsible for the content, and the views expressed do not necessarily represent the positions or policies of the College Board.

1 Overview of Multistage Test/Testing (MST)

A review of the current MST literature traces the origins of MST to research conducted in the 1960s and 1970s. Cleary, Linn, and Rock (1968a) introduced the concept of programmed tests, in which examinees were routed to different measurement sections based on their performance on an initial routing section. The authors further developed this concept by proposing and evaluating various routing strategies (Cleary, Linn, & Rock, 1968b) and comparing alternative types of programmed tests (Linn, Rock, & Cleary, 1969). In parallel, Lord contributed to the theoretical foundation of multistage testing. He explored theoretical aspects of two-stage testing (Lord, 1969, 1971b) and introduced the self-scoring flexilevel test, an early form of tailored testing (Lord, 1971a). Lord (1974) also outlined methods for redesigning a homogeneous test into a multilevel test structure. Building on this work, Waters (1975) examined the utility and validity of tailored testing strategies.

Several studies have provided overviews of MST and discussed its advantages and disadvantages relative to other testing designs. Luecht (2000) characterized MST as computer-adaptive sequential testing (CAST) within a multistage, testlet-based framework, highlighting four key aspects of CAST. Armstrong, Jones, Koppel, and Pashley (2004) presented an overview of MST fundamentals, including design attributes, and outlined mathematical programming techniques for assembling parallel MST forms. Mead (2006) discussed how MST can address certain limitations of computerized adaptive testing (CAT). Kimura (2017) examined CAT from the perspectives of various stakeholders, such as test developers, psychometricians, test takers, educators, and subject matter experts, and suggested that MST may mitigate some of the perceived drawbacks of CAT. Additionally, Stark and Chernyshenko (2006), Hendrickson (2007), and Yamamoto, Shin, and Khorramdel (2018) outlined the benefits of MST in a range of contexts.

2 Implementation of MST

To support the validity and practical utility of MST, numerous studies have demonstrated its application to operational testing programs and evaluated its performance relative to other testing designs, such as paper-and-pencil (P&P) tests and CAT. The following subsections summarize key studies that have examined the implementation and effectiveness of MST in comparison to these alternative formats.

2.1 Applications to Operational Testing Programs

Since the introduction of MST, numerous studies have demonstrated its applicability to operational testing programs. Luecht and Nungester (1998) described how a CAST approach could be used for test construction and administration, using medical licen-

sure examinations as illustrative examples. Luecht, Brumfield, and Breithaupt (2006) provided an overview of MST, including practical considerations for test development, using the Uniform CPA Examination as a case study. Breithaupt, Ariel, and Veldkamp (2005) and Breithaupt and Hare (2007) explored the automated assembly of MST forms for licensing examinations.

In the context of international large-scale assessments, Yamamoto, Shin, and Khorramdel (2018) discussed the advantages of MST over alternative test designs, highlighting its implementation in programs such as the Programme for the International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA). Further, Yamamoto, Khorramdel, and Shin (2018) illustrated how adaptive features can be integrated within the broader design and operational constraints of PIAAC.

2.2 MST, CAT, and P&P

To examine the appropriateness of MST, many studies have compared its performance to that of other testing designs, such as P&P tests and CAT, across various testing contexts. H. Kim and Plake (1993), Patsula (1999), and Reese and Schnipke (1999) compared MST to P&P and CAT formats with respect to the accuracy of ability estimates. Hendrickson (2002) applied a two-stage MST design to the Iowa Tests of Basic Skills (ITBS) and explored how MST could incorporate benefits from both CAT and P&P designs. In more recent work, Sari and Huggins-Manley (2017) manipulated the number of content areas to evaluate the accuracy of ability estimates across CAT and MST. Similarly, K. Wang (2017) used separate item pools for CAT and MST to ensure alignment with content specifications in comparative analyses.

MST, CAT, and P&P formats have also been compared in high-stakes settings, such as licensure testing (Hambleton & Xing, 2006; Jodoin, 2003; Jodoin, Zenisky, & Hambleton, 2006; Xing, 2001). Some studies incorporated testlets into MST designs and investigated performance differences across test formats (Keng, 2008; J. Kim, 2010; Macken-Ruiz, 2008; Schnipke & Reese, 1997). Item pool characteristics were another focus in comparative studies (Jodoin, 2003; Jodoin et al., 2006; Rome, 2017; Xing, 2001), as were IRT model assumptions (Rotou, Patsula, Steffen, & Rizavi, 2007).

Zheng, Nozawa, Gao, and Chang (2012) conducted simulation studies varying panel design configurations, assembly priorities, and routing strategies, comparing MST results to those of CAT and P&P counterparts. Comparative studies have also examined the use of MST and CAT in specific applications such as vertical scaling and large-scale surveys (Beard, 2008; Martin & Lazendic, 2018).

These studies contribute to the growing body of literature evaluating MST in relation to alternative testing designs under a variety of testing conditions and design constraints.

3 MST Studies

The construction of MST panels involves numerous decisions related to both design components and statistical procedures. A growing body of research has explored these aspects to guide the development and implementation of MST.

The first subsection summarizes studies that have examined various options for MST design and statistical methods, such as module structure, routing rules, and scoring procedures. The second subsection highlights studies that proposed novel approaches or methodological innovations within the MST framework.

3.1 Common Study Factors

As noted earlier, the development of an MST panel involves decisions related to both design and statistical components. Design components typically include test length, the overall difficulty level, the number of stages, the number of modules per stage, the difficulty level of each module, and the number of items within each module. Statistical components and procedures encompass calibration methods, routing and scoring strategies, as well as approaches to addressing dimensionality and item local dependence.

The following sections are organized based on the primary focus of the studies reviewed. It is important to note, however, that most studies examined multiple aspects of MST design and statistical procedures simultaneously. For example, while sample size was frequently considered in simulation studies, it was generally not the central focus and therefore does not warrant a separate section. Finally, the studies included in this review are illustrative rather than exhaustive and are intended to provide a representative overview of the MST literature.

3.1.1 Design

Numerous studies have explored the design aspects of MST, focusing on factors such as the number and structure of modules, item types, panel configurations, and item selection strategies.

Cleary et al. (1968b) conducted one of the earliest studies investigating routing strategies by dividing examinees into three or four groups, a decision closely tied to the number of modules available at subsequent stages. Edmonds (2004) and Armstrong and Edmonds (2004) examined the performance of various MST panel configurations with the goal of identifying cost-effective designs that could reduce item development burdens. Similarly, Zenisky (2004) and Zenisky and Hambleton (2004)investigated how MST design features influenced pass-fail decisions and the precision of ability estimates. Svetina, Liaw, Rutkowski, and Rutkowski (2019) studied the allocation of items across modules by varying the number of items per module while keeping overall test length constant.

Several studies have investigated MST designs incorporating polytomously scored items. For example, Chen (2010) explored optimal MST configurations using the generalized partial credit model (GPCM), while J. Kim, Chung, Dodd, and Park (2012) compared panel designs for mixed-format tests in relation to classification accuracy. Park (2015) examined the impact of varying the proportion of polytomously scored items on measurement precision in mixed-format MST panels.

Other studies focused on panel assembly and item selection strategies. K. Wang (2017) evaluated forward and backward methods for setting assembly priorities and recommended the forward method. AlGhamdi (2018) compared 1-3 and 1-3-3 designs, manipulating the range of item difficulties between adjacent modules. Davis and Dodd (2003) examined four item selection methods within a testlet-based MST, two of which incorporated item exposure control mechanisms. Rome (2017) assessed the effectiveness of two item selection approaches under complex content constraints. Park (2013) extended this line of research by introducing statistical constraints at both the module and test levels during panel construction and evaluating their impact on classification accuracy.

Together, these studies demonstrate the diversity of design considerations in MST development and provide empirical evidence to guide best practices in constructing effective MST panels.

3.1.2 Calibration Methods

A number of studies have investigated calibration approaches within MST frameworks. Glas (1988) evaluated item parameter estimates under the Rasch model using both conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation methods. Eggen and Verhelst (2011) further compared the performance of CML and MML approaches in the context of incomplete testing designs, including MST.

More recent work by Cai (2018) and C. Wang, Chen, and Jiang (2020) examined multiple calibration methods–four and three, respectively–and evaluated their relative performance. In the context of multidimensional assessments where each item measures a single dimension, Jewsbury and van Rijin (2020) compared the effectiveness of a single multidimensional IRT model versus multiple unidimensional IRT models for calibration purposes. Extending prior work on calibration within MST, Karatoprak Ersen and Lee (2023) investigated various calibration and linking methods for pretest items.

3.1.3 Routing Strategies

Several studies have examined the effectiveness of different routing strategies within MST frameworks. Armstrong (2002) compared four routing rules that were based on either pre-defined percentages allocated to paths or information functions derived from ability estimates and number-correct (NC) scores. AlGhamdi (2018) evaluated the DPI

and AMI methods with respect to both ability estimation accuracy and classification accuracy. Dallas (2014) compared two routing strategies: one designed to maximize information and another aimed at balancing module exposure.

J. Kim, Chung, Park, and Dodd (2013) investigated routing strategies specifically for MST panels built with polytomously scored items, while Sari and Raborn (2018) evaluated five routing methods, each based on a different type of information. Svetina et al. (2019) also compared five routing approaches and manipulated routing probabilities to study their effects on item exposure and parameter recovery. Additionally, S. Kim and Moses (2014) explored the consequences of misrouting within a two-stage MST design.

3.1.4 Scoring Strategies

A number of studies have explored scoring strategies used in MST, focusing on the accuracy and robustness of different ability estimation methods. Beard (2008) compared the performance of MLE and EAP methods in the context of vertical scaling. Dallas (2014) examined the effects of EAP and NC scoring methods. S. Kim and Moses (2016) evaluated five scoring methods with an emphasis on their robustness to atypical response behaviors. S. Kim, Moses, and Yoo (2015a) and S. Kim, Moses, and Yoo (2015b) conducted simulation studies to compare the performance of seven ability estimation methods. These studies contrasted Bayesian and non-Bayesian estimators and examined differences between NC scoring and pattern scoring approaches.

3.1.5 IRT Assumptions

MST designs typically rely on the assumptions of IRT, including local independence and unidimensionality. A number of studies have investigated the implications of violating these assumptions and explored alternative modeling strategies to address such violations.

- **Local Dependency**

Lu (2010) examined the impact of local item dependence among testlet items on pass-fail classification decisions. Hembry (2014) extended this work by studying mixed-format, testlet-based MSTs under varying conditions of local item dependence, panel design, and routing procedures.

- **Dimensionality**

Y. Zhang (2006) investigated the robustness of a unidimensional IRT model when the unidimensionality assumption was violated, specifically in the context of CAST. X. Wang (2013) evaluated the performance of multidimensional MST designs and compared them to their unidimensional counterparts. J. Zhang (2013) introduced a modified version of the DETECT index for conducting dimensionality analyses within MST and CAT

frameworks. In addition, Jiang (2019) explored factors influencing the accuracy and reliability of subscore estimates.

3.1.6 Other Factors

Researchers have explored a variety of additional issues that influence the design, analysis, and interpretation of MST results.

- **Sample Size and Missing Data**

Chuah, Drasgow, and Luecht (2006) investigated minimum sample-size requirements for stable item-parameter estimation. Cetin-Berber, Sari, and Huggins-Manley (2019) evaluated four imputation techniques for handling missing responses in MST. Jewsbury and van Rijin (2020) also studied missing data, contrasting a single multidimensional IRT model with multiple unidimensional models for tests intended to measure multiple dimensions.

- **Item and Test Characteristics**

Davey and Lee (2011) examined item-position effects in the revised GRE. Gierl, Lai, and Li (2013) examined the performance of Computer Adaptive Testing–Simultaneous Item Bias Test (CATSIB) for detecting differential item functioning (DIF) under balanced and unbalanced MST designs. Sadeghi and Khonbi (2017) provided a practical overview of DIF detection in MST.

- **Speededness and Response Time**

van der Linden, Breithaupt, Chuah, and Zhang (2007) applied a probabilistic response-time model to identify differential speededness and to estimate time-intensity parameters across subtests and examinees.

- **Reliability Estimation**

Y. Zhang, Breithaupt, Tessema, and Chuah (2006) estimated test reliability for a certification exam using two IRT-based procedures. S. Kim and Livingston (2017) assessed the accuracy of a CTT-based reliability estimator for MST scores. Park, Kim, Chung, and Dodd (2017) proposed an analytical approach for computing conditional standard errors of measurement and classification accuracy indices for MST scores.

These studies extend the MST literature beyond core design and estimation concerns, offering insights into practical challenges such as missing data, item drift, DIF, speededness, and reliability estimation.

3.2 Emerging Approaches in MST Research

As MST has gained broader adoption in operational testing programs, the focus of research has gradually shifted from comparing MST to other testing designs toward developing new approaches aimed at enhancing its performance. The first subsection highlights studies that introduced novel methods for assembling MST panels. The second subsection presents research proposing innovations in other areas, including item generation and selection, scoring strategies, and examinee routing.

3.2.1 Advances in MST Form Assembly

Ensuring parallelism across both modules and panels is central to maintaining equivalence and fairness in MST. Early work by Ariel, Veldkamp, and Breithaupt (2006) introduced a modified mixed-integer programming (MIP) approach that simultaneously assembled parallel modules. Subsequent contributions refined top-down and classification-based strategies: Luo and Kim (2018) proposed a route-based top-down method that set design constraints at the test level, and Xiong (2018) organized assembly requirements into six hierarchical levels to optimize panel quality. Furthermore, Z. Wang, Li, and Wothke (2019) utilized the modified NWADH method and developed a heuristic procedure for assembling testlet-based multistage panels.

Researchers have also leveraged test information functions (TIF) as assembly targets. Wu (2001) presented assembly approaches that either maximize the TIF at a single ability point or match targeted TIF values across multiple ability points. Luecht and Burgin (2003) described a strategy for generating feasible TIF targets tailored to MST designs. Armstrong (2005) demonstrated how population ability distributions and item pool characteristics can guide the specification of targets for both the TIF and the test characteristic curve (TCC). Belov and Armstrong (2008) introduced a Monte Carlo-based assembly method that satisfies targeted TIF constraints while enhancing item pool utilization.

When pre-constructed forms proved limiting, adaptive assembly methods emerged. Han and Guo (2013) introduced a method that assembles modules on the basis of the information gap between the current and target TIFs at an interim ability level. Zheng and Chan (2015) developed "on-the-fly assembled multistage adaptive testing" (OMST), which constructs each new module to match the examinee's updated ability estimate. Tay (2015) extended OMST by proposing two additional adaptive variants and a hybrid design. S. Wang, Lin, Chang, and Douglas (2016) presented a hybrid framework in which pre-constructed or on-the-fly MST stages are delivered first, followed by CAT stages for finer adaptation. Zeng (2016) applied OMST to entire test batteries and also explored a hybrid design of MST and CAT. Luo and Wang (2019) advanced this work with dynamic MST, which incorporates both stage- and item-level adaptation. Ma (2020) proposed a passage-based hybrid design that assigns passages to modules while adaptively selecting

the associated items via the MFI rule.

Further efforts have been made to improve item usage in MST. Luecht (2003) introduced a comprehensive test development paradigm aimed at addressing challenges such as item exposure control, content balancing, and reliability. Edwards, Flora, and Thissen (2012) proposed a specialized MST design that achieved a uniform item exposure rate. Xu (2010) extended exposure control techniques developed for CAT and proposed two designs adapted for use in MST. Park, Kim, Chung, and Dodd (2014) developed a method to enhance item pool utilization in constructing mixed-format MSTs based on the GPCM. Yang (2016) and Yang and Reckase (2020) presented strategies for improving pool utilization by assembling parallel modules and panels while maintaining acceptable levels of measurement accuracy.

Other studies have proposed innovations targeting specific testing contexts. Pohl (2013) developed a method for applying MST in longitudinal large-scale studies. C. Wang, Zheng, and Chang (2014) developed a diagnostic MST design and introduced new indices for evaluating item difficulty and test reliability.

Recent work has also extended MST into cognitive diagnostic frameworks. Kaplan (2016) proposed several item selection methods tailored to cognitive diagnostic assessment. the authors made the design feasible by proposing new indices that could measure item difficulty and reliability of test.

3.2.2 Additional Advances in MST Research

Several studies have proposed new approaches to item selection, scoring, and examinee routing within MST. Colvin (2014) applied automatic item generation techniques in the MST context to enhance item pool development. Du, Li, and Chang (2019) integrated response time into item selection procedures within an OMST framework. Han, Dimitrov, and Al-Mashary (2019) introduced new scoring methods and proposed criteria for selecting subsequent modules. Additional work has focused on routing strategies. Weissman, Belov, and Armstrong (2007), Yan (2010), and Han (2020) each proposed and demonstrated novel routing algorithms to guide examinees from one stage to the next. Park et al. (2017) developed an analytic method for estimating the measurement precision of MST scores without relying on simulation.

Appendix A Summary of MST Studies

AlGhamdi, H. M. (2018). *Assessment of multiple-form structure designs of multistage testing using IRT* (Unpublished doctoral dissertation). The University of Denver.

Description: The study utilized 2PL items from the General Aptitude Test (GAT) classical linear testing (CLT) to construct multistage test (MST) forms for performance comparison. Simulation considered 1-3 and 1-3-3 structures, with nine items in the first-stage module and three items in each subsequent modules. Assembly and routing conditions were varied across designs. For each structure, panels were assembled under two conditions–narrow range (NR) and wide range (WR)–based on the range of item difficulty between adjacent modules. Separate panels were assembled to measure verbal (GAT-V) and quantitative (GAT-M) abilities. For routing, the study employed the DPI and AMI methods. The DPI method established cutscores such that 30%, 40%, and 30% of population were routed to the easy, medium, and hard modules, respectively. Results were evaluated in terms of ability estimation accuracy, correlation of ability estimates between CLT and MST, and classification accuracy.

Key Findings: The finings indicated that the accuracy of ability estimates was robust to choices for the three study factors. However, correlations between ability estimates obtained from MST and CLT were higher for the three-stage MST structure than for the two-stage structure, especially when the AMI method was employed for routing.

Key Limitations: The item pool size was limited in size and lacked sufficient numbers of very easy and very difficult items. As a result, none of the MST paths considered in this study were able to satisfy the content specification of the GAT.

Ariel, A., Veldkamp, B. P., Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement*, *30*(3), 204-215.

Description: The study employed a mathematical programming technique to assemble parallel modules simultaneously so that equivalence and fairness could be ensured across different MST forms administered to examinees. Specifically, the study adopted the assembly algorithm proposed by van der Linden and Boekkooi-Timminga (1989) with a modified MIP framework to assemble parallel testlets. For the empirical illustration, a master pool of 1,066 MC items from a high-stakes testing program–calibrated using the 3PL IRT model–was utilized. This example demonstrated how the mathematical programming model approach could be implemented to assemble parallel modules while accommodating practical operational requirements in large-scale testing programs (e.g., content constraints, difficulty of testlets, and module parallelism). The module assembly algorithm was solved using the AIMMS programming language with CPLEX 8.1.

Key Findings: The results demonstrated that the mathematical programming technique in satisfying the principles of parallel module construction. Additionally, the study found that determining the location of a module's maximum information depended on the item pool information.

Key Limitations: Since the study employed a relative target information function as the objective function, the modules were not always able to meet their maximum information values at the target ability points. Incorporating additional theta points as targets may facilitate the construction of more parallel information functions across modules.

Armstrong, R. D. (2002). *Routing rules for multiple-form structures* (ETS Research Report No. 02-08). New Town, PA: Law School Admission Council.

Description: The study compared the performances of four routing rules within a multiple-form structure/testlet (MFS) test design. The four routing rules are: (1) assigning examinees to the module with maximum information based on an ability estimate, (2) classifying examinees into percentile groups specified by the design using an ability estimate, (3) assigning examinees to the module with maximum information based on a number-correct (NC) score, and (4) classifying examinees into percentile groups specified by the design using an NC score. For simulation, item pools from the LSAT were employed to construct a variety of MFS designs. Several hundreds of simulees were generated from both normal and uniform distributions, as well as from percentile-based groups of those distributions. Performances of the routing rules were evaluated in terms of score accuracy and classification accuracy.

Key Findings: With respect to RMSE, Rule 1 outperformed the other rules, whereas Rule 4 performed better than Rule 1 in terms of classification accuracy. However, the observed differences were minimal, suggesting comparable scoring accuracy across the four routing rules. The study further noted that routing based on NC scoring could offer practical advantages, including ease of interpretation and the ability to predict examinee distribution across module/testlet. Additionally, Rule 4 required no assumptions regarding TIFs. Both Rule 3 and Rule 4 demonstrated improvements in classification accuracy without any loss in overall scoring precision.

Key Limitations: Since the routing rules in this study were based on specific assumptions, it is important to further investigate the impact of those assumptions on performance outcomes. Future research should also address additional issues such as testlet effect, person-fit statistics, and connection between routing and scoring methods. Moreover, it remains an open question whether the conclusions drawn from this study could be generalized to operational contexts.

Armstrong, R. D. (2005). *A method to determine targets for multi-stage adaptive tests.*

Description: This study proposed a method for determining target TIF and TCC values to assemble MST forms, based on an item pool and the population ability distribution. To establish targets for each bin (i.e., module), the method defined an objective function that accounted for randomization, item exposure rates, and the amount of information in selecting items. Item selection was conducted under an omniscient testing framework where examinees' true abilities were assumed to be known prior to the administration. For each bin, the targets for TIF and TCC were computed as the sum of observed item information functions and characteristic curves weighted by the probability of each examinee visiting the bin. Simulations employed three distinct LSAT item pools and constructed eight non-overlapping MST designs. To evaluate performance, the study computed the correlation between scale scores derived from true abilities and those derived from estimated abilities, the unconditional standard error of scale scores, and the expected scale scores.

Key Findings: The study indicated that the TIF and TCC targets derived using the proposed method successfully achieved the intended properties for the eight MST designs. Moreover, the method provided the flexibility in specifying bins to target particular percentiles of the population, allowing control over scoring accuracy at designated ability levels.

Key Limitations: The procedures for determining targets were relatively lengthy, making it challenging to modify the targets once a test based on them had been administered operationally.

Armstrong, R. D., Edmonds, J. (2004). *A study of multiple stage adaptive test designs.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Description: The study compared the performance of several MST designs, aiming to identify a design that could reduce item development costs. The study conducted simulation by varying the number of stages (3-stage and 4-stage), the number of testlets per bin (5 and 6), and rules for partitioning bins at the 4th stage (even splits versus 20% and 13% allocations in the outer bins). Form assembly utilized an operational item pool from the P&P LSAT test that consisted of 1,336 items calibrated using the 3PL IRT model. Items for each bin were selected to achieve target TIF and TCC values. Scoring and routing were based on NC scores. Assembly problems were formulated as MIP models and solved using CPLEX, with additional functions implemented in AMPL. Results were evaluated in terms of (a) scoring accuracy, (b) design simplicity to facilitate review, and (c) efficiency in item pool usage.

Key Findings: The results indicated that evenly splitting the test-taking population across modules at the fourth stage yielded more efficient item pool usage, whereas the 13% allocation rule produced higher scoring reliability than the alternative rules. The reliability of the 5-testlet designs was lower than that of the 6-testlet designs; however, with other conditions being fixed, the 5-testlet design produced three more forms per assembly than the 6-testlet design did. With respect to design simplicity, item pool usage, and scoring reliability, the 4-stage design did not outperform the 3-stage design, contradicting the conventional expectation that additional stages/adaptation enhance scoring precision. The study results further suggested that an item pool should be composed of items whose characteristics are well-aligned with the tests of interest.

Key Limitations: The study utilized an operational item pool originally developed for linear testing rather than for adaptive testing; therefore, the main findings may be specific to the characteristics of the items within the pool used for assembling linear forms.

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.

Description: This article provided an overview of the fundamentals of MST–also referred to as multiple-form structure (MFS) in this paper. The authors introduce MFS as a method that combines the benefits of traditional P&P with the adaptability of CATs. The article discussed key attributes of the MFS design (MFSD) including modules/bins, bin targets (e.g., TIF and TCC), the number of items per bin, the number of stages, and routing rules, as well as practical constraints such as content, answer key count, word count, topic coverage, and diversity usage. The article also outlined mathematical programming techniques for assembling parallel MST forms. A computerized LSAT was presented as an illustrative example to demonstrate the attributes of the MFSD. Additionally, a study was conducted to compare the performances of P&P, CAT, and a 6-stage 1-1-2-3-3-4 MST versions of the test. The study utilized an item pool provided by LSAT and implemented the shadow assembly approach proposed by van der Linden and Reese (1998) for CAT. Constraints of the MSFDs were similar to those of the P&P LSAT sections. For MSFDs, NC scores were employed for both scoring and routing. The performances of the three approaches were evaluated in terms of the accuracy of ability estimates (i.e., RMSE) and the correlation between true and observed scale scores.
Key Findings: The results indicated that, compared to CAT, MST can be considered as feasible approach, as it provides better control over item exposure parameters while achieving comparable reliability.
Key Limitations: The MFS assembly method did not guarantee an optimal solution to the mathematical programming problems. Moreover, the authors recommended future research that investigates various factors influencing item pool usage, suggesting that the present study did not fully explore all the factors affecting pool usage.

Beard, J. J. (2008). *An investigation of vertical scaling with item response theory using a multistage testing framework* (Unpublished doctoral dissertation). The University of Iowa.

Description: This study investigated the effects of different testing frameworks and statistical estimators on the construction of a vertical scale across three years using IRT scores. A simulation study was conducted considering two testing frameworks (MST and single form testing [SFT]) and two statistical estimators (MLE and EAP). For MST, a base form was constructed for the initial year, with easy and hard forms developed for the first and second follow-up years. For SFT, three fixed forms were constructed, each targeting a different year through varying difficulty level. In both MST and SFT, test forms included common items across administrations and within each administration. The simulation involved 5,000 examinees, each with three abilities generated from a correlated multivariate normal distribution. For MST, routing was based on examinees' true abilities. At the end of testing, the study fixed item parameters and estimated final abilities (MLE or EAP) using the multi-group and concurrent calibration method. Results were evaluated in terms of ability recovery, recovery in growth scores, and the separation of score distributions across the three years.

Key Findings: The results indicated that employing the MST framework in conjunction with the EAP method improved both ability estimation and growth score recovery. However, for both MST and SFT, the separation of score distributions was better preserved when using the MLE method, whereas EAP produced some irregularities in the horizontal distances between distributions.

Key Limitations: For MST, routing was based on true abilities, a condition that can not be attainable in operational testing. Moreover, since the multi-group concurrent calibration method treated the group of examinees taking the easy forms as the reference group, potential bias may have affected the ability estimates. Additionally, abilities were estimated by fixing item parameters; if item parameters were simultaneously estimated with abilities, the irregular patterns observed in EAP ability estimates might be mitigated.

Belov, D. I., Armstrong, R. D. (2008). A monte carlo approach to the design, assembly, and evaluation of multistage adpative tests. *Applied Psychological Measurement*, *32*(2), 119-137.

Description: The study proposed a Monte Carlo (MC) based method for assembling MST forms. The study used an LSAT item pool that consisted of 158 passages and 1,470 items calibrated using the 3PL IRT model. The study considered a 1-1-2-3 panel design and applied the MC method with 12 constraints (see the original article for details) to assemble each module. Additionally, the study examined three approaches to enhance the likelihood of meeting target TIFs and to improve overall pool utilization. The three approaches were (1) adding items with the highest usage frequency to the existing item pool as new items, (2) shifting the ability distribution to better align with the pool information function, and (3) reducing the number of linear forms in the MC algorithm. Results were evaluated in terms of the number of overlapping/nonoverlapping items among forms, item-level usage frequency, and the pool utilization index, defined as the ratio of the number of distinct items used to the total number of items in the pool).
Key Findings: Among the 12 constraints considered in the study, satisfying TIFs proved to be the most challenging. All three approaches enhanced the alignment with target TIFs and improved pool utilization. The MC algorithm demonstrated sufficient flexibility to accommodate additional constraints and to support different IRT models.
Key Limitations: The study considered constraints applied only to path; future research could explore the incorporation of additional, potentially nonlinear constraints.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., Moser, U. (2019a). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement*, *56*(1), 121-146.

Description: As a means to better align students' abilities with item difficulties, the study introduced targeted multistage (TMST) calibration and compared the efficiency of item calibration across different designs. TMST combines features of the MST design and the targeted calibration (TC) design, in which students are assigned booklets based on their background variables. In TMST, students are routed to either an easy or a difficult first-stage module according to background information, after which MST procedures are applied. For simulation, the study examined four calibration designs: (1) TC, (2) TMST with five difficulty categories for MST, (3) TMST with four difficulty categories for MST, and (4) random assignment. To model scenarios with limited prior knowledge of item difficulty, the study also manipulated the correlation between item difficulty and item order, with values of 0.4, 0.6, or 1.0 (the optimal condition). Test lengths were fixed to 20 items for all designs. For each study condition, 20,000 data sets were generated, each with 2,600 simulees from either $N(0, 1)$ or $N(0.8, 1)$. Performance was evaluated in terms of the item difficulty distribution per booklet/module, the accuracy of item difficulty estimates, and the average number of observations per item.

Key Findings: Under the optimal conditions, the two TMST designs yielded more accurate estimates of item difficulty parameters than the other designs, particularly for items at the lower and upper ends of the difficulty spectrum. However, when prior knowledge about item difficulty was limited, the findings recommended the use of the TC design.

Key Limitations: Since the study was conducted under limited conditions (i.e., using the Rasch model and a small sample size), the generalizability of the findings should also be restricted. Moreover, the study focused only on the alignment between students' abilities and item difficulties. Future study should explore factors such as the strength of links between modules (e.g., the number of common items) as a potential way to enhance calibration efficiency.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., Moser, U. (2019b). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontier in Education*, *4:1*.

Description: The study investigated whether targeted MST (TMST) designs could enhance measurement efficiency relative to alternative designs. In TMST, the first-stage module is selected based on an examinee's ability-related background variables. For comparison, the study also evaluated a linear design, a targeted design with three forms, a 1-3-3 MST design, and a 3-3-3 TMST design. Across all designs, the test length was fixed at 30 items. For the MST and TMST designs, the length of the first-stage modules was varied to represent 1/5, 1/4, 1/3, or 1/2 of the total test length. For each study condition, the study generated 1,000 data sets, each with 30,000 simulees from a combination of three population distributions. The degree of overlap among the distributions were systematically manipulated (narrow, medium, and wide). Design performance was evaluated with respect to the accuracy of ability estimates and efficiency gains relative to the MST and TMST designs. Additionally, the study conducted the analysis of variance to assess the effects of the study factors. The alignment between true ability and module difficulty was also examined in terms of the percentages of correctly assigned, slightly incorrectly-assigned, and heavily incorrectly-assigned examinees.
Key Findings: Compared to the targeted design, both the MST and TMST designs improved the measurement efficiency. When the ability range of the target population was wide and the ability-related background variable was reliable, the TMST design proved to be effective. However, when the ability range was narrow and the ability-related background variable lacked reliability, measurement efficiency was comparable across the MST and TMST designs. Furthermore, the length of the starting modules had minimal impact on measurement efficiency within the TMST designs.
Key Limitations: The study was conducted under limited conditions (e.g., Rasch items, fixed test length, and specific MST/TMST designs). Future research should extend these findings by incorporating alternative conditions (e.g., different IRT models, longer test lengths) to examine the robustness and generalizability of the results.

Breithaupt, K., Ariel, A., Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing*, *5*(3), 319-330.

Description: This study presented design principles for developing the computerized version of the Uniform CPA examination and discussed their practical implementation. An illustrative example was provided to demonstrate practical solutions for challenges in automatically assembling test forms under conditions of limited item pools and strict content security. The example considered a 1-2-2 MST design, which was adopted for the computerized version of the Uniform CPA examination. In assembling testlets/modules, the study highlighted four design-related issues that required sequential decisions: (1) defining the assembly objective, (2) determining the location and type of IRT information targets, (3) specifying the number of possible testlets, and (4) setting the stringency of content constraints. Once these decisions were established, constraints and variables were formulated as relative optimization functions for each testlet. MIP was used to solve the optimization functions and to assemble a large number of parallel testlets. Final panels were reviewed by the psychometric and content experts. When it was not possible to generate testlets that met all constraints (i.e., the MIP problem was "infeasible"), the issue was resolved by adjusting an initial constraint (e.g., the number of items in one content area) to the minimum acceptable value within a desirable range.
Key Findings: The study demonstrated that, for the CPA examination, diverse constraints in test form construction could be addressed through the formulation of optimization functions. The example further illustrated that testlets could be assembled by solving those functions while preserving desirable psychometric properties. The study suggested that the proposed design principles and assembly procedures could be applied to other testing programs, provided that their psychometric properties are maintained. Furthermore, there is a trade-off: while strict content constraints help ensure parallel testlets, they also reduce the number of testlets that can be generated.
Key Limitations: When the number of constraints and variables is large, solving the optimization functions using MIP can be computationally demanding or may lead to infeasibility. This issue was particularly pronounced when the quality of item pool was not well understood or when items were scarce in specific content areas.

Breithaupt, K., Hare, D. R. (2007). Automated simultaneous assembly of multi-stage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5-20.

Description: To provide evidence of the feasibility of MIP for assembling MST testlets and modules, the study applied the algorithm to the construction of a computer-adaptive MST (ca-MST) for the AICPA licensing exam. Using the MIP method, the study aimed to identify an optimal solution in which all testlets could be assembled simultaneously while satisfying all constraints related to maximizing TIF at targeted abilities, test length (i.e., the number of items per module), content and skill specifications, and item enemy rules. Assembled modules were then combined into panels using linear programming. Panel-specific constraints included the number of enemy items, the number and placement of pretest items, and the maximum allowable number of module reuse. The performance of the MIP method was evaluated in terms of module and item exposure rates. Additionally, an optimized panel was constructed as a baseline satisfying the minimum possible exposure for any testlet.

Key Findings: The MIP method successfully facilitated the assembly of the ca-MST design for the CPA exam. Subsequent studies further demonstrated the applicability of the method for constructing tests with varying constraints on psychometric models and content composition. Moreover, the approach allowed for the simultaneous construction of multiple parallel forms of linear tests as well as optimal nonparallel forms such as modules across different stages.

Key Limitations: The study focused exclusively on the testlest, without considering the accounting performance simulation tasks. Furthermore, the discussion on item exposure should be limited to individual items within one example administration; longer term inventory management, which depends on subpool selection and an optimized schedule for item writing and retirement, was beyond the scope of the study. Additionally, the study noted that conditional exposures, such as those within a high-ability candidate group, remained unexplored.

Cai, L. (2018). *An investigation of item calibration approaches in multistage testing* (Unpublished doctoral dissertation). The University of Nebraska.

Description: The study evaluated the performances of four item calibration methods within the context of MST. The methods were: (a) separate calibration with linking (SCL), (b) fixed common item parameter calibration (FCIP), (c) concurrent calibration with a single group (CCSG), and (d) concurrent calibration with the multiple-group procedure (CCMG). The study implemented a 1-3 MST design, with the total test length fixed at 35 items. For simulation, the study varied the length of the routing test (10, 15, and 20 items), sample size (4,200 and 12,000 from $N(0, 1)$), and the routing rule. Two routing rules were employed: approximate maximum information (AMI) and defined population intervals (DPI). Both interim and final ability estimates were obtained using the EAP estimator. At the end of testing, examinees were classified into one of four performance levels. The four calibration methods were applied to estimate item parameters, which were subsequently used to re-estimate final abilities. For each study condition, the study performed 100 replications, and results were evaluated in terms of item parameter estimation accuracy, ability estimation accuracy, and classification accuracy. In addition to the simulation study, the study also used empirical data to compare pre-equated and post-equated item parameter estimates.
Key Findings: The performances of FCIP and CCSG were comparable and reasonably good, whereas CCMG demonstrated notably poor performance. The performances of SCL were even worse, to the extent that its feasibility was questioned under the study conditions. Ability estimation improved as the number of items in the routing module increased. While AMI and DPI yielded similar accuracy in ability estimation, AMI produced higher classification accuracy compared to DPI.
Key Limitations: Since the simulation study was conducted under limited conditions, the findings should not be generalized beyond those specific settings. Future study could extend the investigation by applying the calibration methods to different item types, alternative population ability distributions, and additional elements involved in constructing MST panels.

Cetin-Berber, D. D., Sari, H. I., Huggins-Manley, A. C. (2019). Imputation methods to deal with missing responses in computerized adaptive multistage testing. *Educational and Psychological Measurement*, *79*(3), 495-511.

Description: The study evaluated the performance of various imputation methods for handling missing data within the MST framework. For simulation, the study varied conditions for MST designs (1-3, 1-2-2, 1-2-3, and 1-3-3), test length (30 and 60 items), percentage of missingness (5%, 15%, 30%, and 50%), imputation method, and type of missingness. Four imputation methods were compared: (1) person mean imputation (PMI), which computed the mean score of the observed responses of a person, (2) predicted probability imputation (PPI), which adapted a logistic regression approach, (3) full information maximum likelihood (FIML), which utilizes all available cases to estimate parameters, and (4) incorrect, which treated missing responses as incorrect. Three types of missingness were considered: MCAR, MAR, and MNAR. As a baseline condition, the study generated a complete data without any missing responses. For each MST design, item parameters from a previous study were used to construct three nonoverlapping parallel panels consisting of easy, medium, and hard modules with maximized information at θ = -1, 0, and 1, respectively. For each study condition, 100 datasets were generated, each with 900 simulees from $N(0, 1)$. Examinees were routed to the module providing maximum information at their interim ability estimates. Both interim and final abilities were estimated using the EAP method. Results were evaluated with respect to the recovery of ability estimates (bias and RMSE) and correlation between true and estimated abilities.

Key Findings: Across all types of missingness, the PMI, PPI, and FIML methods outperformed the incorrect method. At the 5% missingness, these three methods produced results comparable to the baseline condition, and ability estimates were also acceptable up to the 30% missingness. However, at the 50% missingness, all methods yielded biased ability estimates. Among them, FIML demonstrated the greatest robustness, showing the most consistent results across study factors and the lowest RMSE. The performance of PMI was generally comparable to that of the other imputation methods. In contrast, test length and MST design seemed to have minimal impact on outcomes.

Key Limitations: The study used the single imputation only; future research could explore multiple imputation techniques for handling missing values. In addition, future research should also examine missingness arising from time limits and consider issues from test-takers' perspectives.

Chen, L.-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: The study employed polytomously-scored items modeled with the GPCM to identify optimal designs for MST. To assemble MST panels, the study utilized a pool of 273 items from a national test. The study considered eight panel designs (1-2, 1-3, 1-4, 1-2-2, 1-2-3, 1-2-4, 1-3-3, and 1-3-4). For each design, the study varied routing test length (short and long) and total test length (short and long). Items were selected to meet target TIFs and content balance requirements. Both interim and final abilities were estimated using the MLE method. Examinees were routed to modules providing maximum information given their estimated ability levels. For each simulation condition, the study conducted 10 replications, each with 1,000 examinees from $N(0, 1)$. Results were evaluated with respect to the accuracy of ability estimates, item exposure rates, item overlap, and item pool utilization.

Key Findings: The study found results to be highly consistent across all test designs. The impact of the MST structure on the accuracy of ability estimates was minimal when items were polytomously scored under the GPCM. However, the results showed that non-convergent cases in ability estimation were more frequently observed when routing tests were shorter.

Key Limitations: Since the study factors and conditions were limited, it should be careful when generalizing the results to other contexts. To validate the findings, future research should expand the scope by considering additional variables such as item pool characteristics, test assembly methods, ability distributions, the difficulty of the first-stage module, and item types.

Chuah, S. C., 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.

Description: The study examined the required sample size for achieving adequate estimation of IRT item parameters within the context of CAST (Luecht, 2000; Luecht & Nungester, 1998, 2000). For simulation, the study used a pool of 450 items from a P&P certification exam. First, pretest data sets were generated by simulating a sparse response matrix. The first 50 items were designated as anchor items and administered to the first subgroup. Each subsequent subgroup was then administered to 45 unique items along with 5 items from the anchor set. For subgroup sample size (i.e., sample size per item), three levels were used: 300, 500, and 1,000 from $N(0, 1)$. Items were calibrated using the 3PL model. Then, for each of the four sets of item parameter estimates (three from the pretest samples and one from the item pool), the study assembled two 1-3-3 parallel panels, with 20 items per module. For each panel, 5,000 simulees were generated from $N(0, 1)$ and abilities were estimated using the MLE method. Routing decisions were based on NC scores to ensure balanced module exposure across examinees. Results were evaluated with respect to item information, the accuracy of ability estimates, and the accuracy classification into master and nonmaster categories.
Key Findings: Regarding ability estimation and classification accuracy, a sample size of 300 per item was sufficient. However, this sample size was inadequate for accurate estimation of item parameters, as discrimination parameters tended to be overestimated.
Key Limitations: The study assumed that all simulees were fully motivated and possessed identical ability parameters in both the pretesting and operational testing contexts. Additionally, it was assumed that the ability distributions of simulees in the pretest samples was the same as those of examinees completing the CAST, an assumption that may not hold in actual testing situations.

Cleary, T. A., Linn, R. L., Rock, D. A. (1968a). An exploratory study of programmed tests. *Educational and Psychological Measurement*, *28*, 345-360.

Description: The study described programmed tests that could reduce testing time while maintaining an acceptable level of reliability. This study considered programmed tests that consisted of two sections: (1) a routing section, which could direct examinees to an appropriate subsequent measurement section, and (2) a measurement section, with item difficulties were tailored to the examinees' ability levels. To develop the routing section with 20-23 items, the study considered four methods, two-stage, broad-range, group-discriminating, and sequential (please see the original article for details). For the measurement section, the study constructed four tests of 20 items each, corresponding to four distinct ability levels, which were consistently used across all routing methods. Following the routing test, examinees were assigned to one of the four measurement sections, with group sizes approximately equal. To evaluate the applicability of the programmed tests, the study utilized real data comprising responses from 4,887 students to 190 items. Performance was assessed based on correlations between NC scores on the entire 190 items and scores on the programmed tests, split-half and Kuder-Richardson formula 21 reliabilities for the 190 items, and part-whole correlations between the best 40-item subtest and the total test and between the best 42-item subtest and the total test.

Key Findings: Correlations between the original test scores and scores on the programmed tests were very high. Among the four methods for developing the routing section, the sequential method produced the best results. However, the study noted that comparable results could also be achieved using fewer items if conventional tests were constructed solely with the statistically optimal items.

Key Limitations: The study applied the programmed tests using the existing data. To obtain more conclusive evidence regarding their performance, the programmed tests should be administered to examinees in operational settings. Additionally, the effectiveness of programmed tests should be evaluated against external criteria that are experimentally independent.

Cleary, T. A., Linn, R. L., Rock, D. A. (1968b). Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, *5*(3), 183-187.

Description: The study aimed to find the most efficient routing strategy for programmed tests, in which subjects are routed to a set of items tailored to their performance levels. The study proposed two sequential routing strategies: (1) a four-group method and (2) a three-group method. In the four-group method, 23 items were administered in the routing section, subjects were assigned to one of four ability groups; in the three-group approach, 20 items were administered, and subjects were assigned to one of three ability groups. The second-stage measurement tests consisted of 20 items that had the highest within-group point-biserial correlation with total test scores, excluding items from the routing section. ETS response data for 190 items answered by 4,885 subjects were used for test assembly. Subjects were randomly assigned either an original or a cross-validation group and completed the programmed tests. Additionally, the study constructed best short tests of varying lengths (5 to 50 in increments of 5). At the end of testing, total scores were predicted by scaling measurement test scores using linear regression weights. Results were evaluated based on correlations between the total test scores and the final predicted total scores.

Key Findings: For both groups, correlations obtained from the programmed tests were comparable to the reliability estimates for the full 190-item test. Relative to the "best" 50-item short test, the three-group method gave a higher correlation. Moreover, a traditional test would require approximately 35% more items than the programmed test employed for the three-group method.

Key Limitations: Since the items in the item pool were developed for measurement rather than routing purposes, the generalizability of the study results should be limited to pools with similar characteristics. Furthermore, the routing strategies examined in this study should not be regarded as optimal. The study recommended that future research consider other factors, such as item format, number of items, that may affect the validity and reliability of the test.

Colvin, K. F. (2014). *Effect of automatic item generation on ability estimates in a multistage test* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Description: This dissertation implemented automatic item generation (AIG) within a MST framework and examined its effects on ability estimation. Following Sinharay and Johnson (2008), item clones were generated by introducing variations to the discrimination and difficulty parameters of parent items. The study applied three levels of variation to the difficulty parameters: small, moderate, and large. For discrimination parameters, random values from a uniform distribution were added to the logarithms of the parent items' parameters. Responses were generated based on the cloned item parameters. Parent items were calibrated using 3PL model. For the percentage of item clones, the study considered 33%, 50%, and 100%. For simulation, 1-2, 1-3, 1-2-4, and 1-3-5 MST designs were implemented with a total of 36 items. For each design, in addition to the baseline condition without cloned items, the study varied both the percentage of clones items and the degree of difficulty variation. For each study condition, 100 data sets were generated, each with 1,000 simulees sampled from 61 ability points raning from -3 to 3 in increments of 0.1. Routing was performed based on NC scores. Final ability estimates were obtained from TCCs computed using parent item parameters, with separate relationships maintained for each design and path. Results were evaluated in terms of the accuracy in ability estimates (bias and SEE) and compared against those for the baseline conditions.

Key Findings: Ability estimates became less accurate as the proportion of cloned items increased and/or the degree of variation from parent items grew. Nevertheless, the study identified certain conditions under which AIG could be implemented without compromising the integrity of the tests.

Key Limitations: Future research could consider evaluation criteria beyond the recovery of ability estimates. Additionally, the study could be expanded to other contexts, such as for tests with different objectives (e.g.,classification) or pattern scoring. Furthermore, the study assumed that item parameters for non-cloned items were error-free, which is unrealistic; as a result, the observed impact of AIG may have been smaller than it would be under more realistic conditions.

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

Description: This study examined the effects of routing and scoring strategies within the framework of ca-MST. For simulation, the study considered two main factors: MST design configurations and item pool characteristics. For design, the study considered three configurations (1-3, 1-2-3, and 1-2-3-4) and two module lengthes (10 and 20 items). For item pool characteristics, the study used three levels of average item difficulty (-1, 0, and 1) and two levels of average item discrimination (.6 and 1). For scoring, two methods were considered: EAP estimation and NC scoring. For routing, two criteria were applied: maximizing module information and balancing module exposure. When the NC scoring method was used, final ability estimates were derived from TCCs. The performance of the four scoring and routing strategies was compared in terms of bias and RMSE of final ability estimates.

Key Findings: The routing strategy based on maximum module information generally outperformed the strategy focused on balanced exposure of modules. Between the two scoring methods, ability estimates obtained with the EAP method were slightly more accurate than those obtained from using NC scores, although the differences were not large. The RMSE decreased as both module length and the number of stages increased. Automatic test assembly performed best when the average item difficulty of the pool was 0, yielding smaller RMSEs compared to other difficulty levels.

Key Limitations: For some conditions, using the default BILOG commands led to convergence issues, and the resulting biases (though not RMSE) were underestimated.

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 No. RR-11-26). Princeton, NJ: Educational Testing Service.

Description: The study examined whether item position effects existed in GRE and evaluated their potential impact on the revised GRE (rGRE) under an MST framework. The study implemented three data collection designs by embedding a pretesting section within each test-taker's operational test. In the first design linear forms were constructed for the quantitative (28 items) and verbal (30 items) sections. Thirteen (quantitative) and seven (verbal) scrambled forms with varying item orders were created, and one scrambled form was randomly assigned to each examinee as a variable section attached to the GRE. The second design arranged quantitative (verbal) items into three sets of 28 (30) items, with each set scrambled in 13 (7) ways with items in various locations; these scrambled sets were then randomly administered to examinees. For the third design, the study applied an MST structure to the pretesting section, using a 1-5-5-5 configuration with 8-8-8-4 items for the quantitative section and 8-10-8-4 items for the verbal section. Item parameter estimates from the first and second designs were linked using the operational GRE section. Since the MST-based rGRE would rely on the 2PL IRT model, pretest items were calibrated under this model, while operational GRE scores continued to be estimated using the 3PL model. The study compared item difficulties across different item orderings and evaluated position effect in terms of the accuracy of predicted scores based on item parameter estimates across scrambled orderings.
Key Findings: The study identified position effects in the GRE data, mostly driven by test speededness. Accordingly, the study recommended that time limits should be generously set, particularly for high-performing examinees. The findings also indicated that using scrambled forms with varied item orders for pretesting items can help mitigate position effects to some extent.
Key Limitations: The study did not allow the administration of tests containing the new item types that the rGRE will introduce in MST.

Davis, L. L., Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the mcat. *Applied Psychological Measurement*, *27*(5), 335-356.

Description: The current study compared the performance of four item selection methods for testlets, including two methods with item exposure control and two methods without. For exposure-controlled methods, the study examined a modified within .10 logits randomization procedure (Lunz & Stahl, 1998) and a CAST. For methods without exposure control, maximum information and random item selection methods were considered. The study used the MCAT data, which included 149 passages calibrated using the PCM. For CAST, the study manually assembled eight 1-3-3 panels based on a test specification, and 1,000 subjects from $N(0, 1)$ were randomly assigned to one of the eight panels. Both interim and final ability estimates were obtained using MLE, and subjects were routed to the module that provided the most information at the interim ability estimates. The study also conducted CAT simulation. Performance across the four item selection methods was evaluated in terms of measurement precision (SE, defined as the inverse of square root of TIF), the correlation between true and estimated abilities, accuracy of ability estimates (bias, RMSE, and MAD), and test security (item exposure rate and item overlap).

Key Findings: In terms of test security and measurement precision, the two methods with item exposure control outperformed those without the control. Compared to the other three methods, CAST offered an additional advantage: by constructing test forms in advance of administration, it enabled stronger quality control over the psychometric and measurement properties of test forms.

Key Limitations: Since item exposure rates could vary depending on the size and structure of the item pool, the findings should not be generalized to other settings with different pool characteristics. Future research should therefore examine item pools with varying characteristics. Also, future research investigate alternative CAST structures and their effects on test security and measurement precision.

Du, Y., Li, A., Chang, H.-H. (2019). Utilizing response time in on-the-fly multistage adaptive testing. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 107-117). Springer Nature Switzerland AG: Springer.

Description: This study proposed incorporating response time (RT) information into item selection algorithms within the framework of OMST. The study examined three RT-based methods: (1) a-stratification with b-blocking and time (ASBT), (2) maximum information with beta matching (MIB), and (3) generalized maximum information with time (GMIT). The performance of the RT-based methods was compared with that of two approaches that do not account for RT: a-stratification with b-blocking (ASB) method and the MFI method. For simulation, the study employed a three-stage panel design, with 15 items per stage. To mitigate test overlap, 20 parallel modules were pre-assembled for the first stage. Both stratified and unstratified item banks were considered during item selection. For each study condition, the study generated 1,000 examinees and conducted 50 replications. The performance of the item selection methods was evaluated with respect to accuracy of ability estimation, test efficiency and stability, test security, and item bank utilization.

Key Findings: Overall, testing times for all RT-based methods were, on average, shorter and exhibited smaller SDs compared to those for the conventional methods, improving both test efficiency and stability. Among the three RT methods, GMIT produced the most accurate ability estimates and the most efficient tests relative to ASBT and MIB. However, GMIT demonstrated the lowest item bank usage efficiency, with higher item overlap and exposure rates than ASBT and MIB. Additionally, when the item bank was stratified, test security and item bank utilization generally deteriorated compared to the unstratified condition.

Key Limitations: The conditions of items, item bank, test designs, and examinees were limited; and the conclusions should not be overly generalized to other settings.

Edmonds, J. J. (2004). *The evaluation of multiple stage adaptive test designs* (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey.

Description: This study evaluated several MST designs with respect to the cost of test development using discrete and set-based items from the P&P LSAT. In this study, a discrete item was a dichotomously scored item whose stimulus and question could be treated as a single unit, whereas a set-based item shared a stimulus across multiple dichotomously scored items. Examinees were generated either from a known distribution (e.g., $N(0, 1)$) or from fixed ability points with replication. For the MST designs, the study varied the percentage of bins at the final stage, the parameters for creating bin targets, and the number and position of bins. Three conditions were considered for the final-stage bin percentages: (1) even splits, (2) 20% in each outer bin, and (3) 13% in each outer bin. Bin targets were created following the method described in Armstrong and Roussos (2002). For the number and position of bins, the study assigned two testlets in the first stage, with a maximum of two testlets for subsequent stages. For discrete items, the study considered 3- or 4-stage designs with 5 or 6 testlets per path; for set-based items, the study used 3- or 4-stage designs with 4 or 5 testlets per path. Results were evaluated in terms of design simplicity, efficiency of item pool usage (e.g., maximum number of non-overlapping tests), and scoring accuracy (e.g., reliability and conditional SE of scale scores). MST panels were assembled using MIP algorithms solved with CPLEX (ILOG, 2002).

Key Findings: Regarding item pool usage and design simplicity, the even-split rule performed best for discrete items, while for set-based items, both the even-split rule and the 20% outer-bin rule outperformed the 13% rule. The results indicated that using fewer testlets per panel reduced scoring accuracy, but allowed assembling more forms. Additionally, 3-stage designs outperformed 4-stage designs. Based on item pool usage and scoring accuracy, the study recommended 3-stage designs with two testlets per stage for discrete items, and for set-based items, 3-stage designs with one testlet in the middle stage and two testlets in the remaining stages. To maximize scoring accuracy for LSAT, the results suggested MST designs with three levels at the second stage.

Key Limitations: Since the item pool used in the simulation study were developed for measurement purposes rather than routing, the findings may not be directly applicable to settings with different item pool characteristics. Therefore, future research should examine item pools with varying characteristics to enhance the generalizability of the results. Furthermore, further studies should investigate the effects of MST designs across diverse item pools and consider the inclusion of new item types in future test administrations.

Edwards, M. C., Flora, D. B., Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education*, *25*(2), 118-141.

Description: The study described a specialized version of the MFS (MST) with uniform item exposure rate (uMFS). To assign items to uMFS modules as optimally as possible, the study employed a threshold accepting algorithm (Dueck & Scheuer, 1990) with objective functions achieving four goals: (1) minimizing the average error variance of estimated abilities given summed score, (2) achieving uniform error variance across a range of scores, (3) constructing multiple routing blocks that were randomly equivalent, and (4) selecting cutscores for routing examinees to ensure uniform item exposure. Using 63-item and 50-item subtests from an existing exam, the study simulated linear and uMFS forms of equivalent and shorter lengths (48 and 36 items). For uMFS, the study considered 1-3-3 and 1-4-4 designs and used the aforementioned algorithm to assemble modules. Results were evaluated in terms of item exposure rate, reliability, conditional SE (i.e., SE($\theta|x$)), and block information functions.

Key Findings: Compared to the 63-item linear test, the 48-item uMFS test achieved equivalent reliability while providing better measurement at the tails of the score scale and more uniformly distributed SEs across the score scale. Similar results were observed for the 36-item uMFS relative to the 50-item linear test. The study demonstrated that uMFS could capture the advantages of CAT while mitigating disadvantages of CAT, such as underutilization of the item pool, limited control over item exposure, and content imbalance, suggesting that uMFS could serve as an effective alternative to CAT.

Key Limitations: The currently designed algorithm does not actively incorporate pre-existing testlets into the item bank. Moreover, since relative efficiency can vary depending on the specific items in the item bank, it would be difficult to generalize the findings to other pools. The authors also noted that any MFS-based adaptive test, including uMFS, is generally less efficient in terms of reliability than an item-by-item CAT. Similarly, implementing exposure control–a core feature of the uMFS–reduces reliability efficiency compared to tests without such controls.

Eggen, T. J. H. M., Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica: International Journal of Methodology and Experimental Psychology*, *32*(1), 107-132.

Description: The study examined the validity of item parameter estimation in incomplete testing designs. A simulation study was conducted varying conditions across both incomplete design and estimation procedure. For estimation, two methods were considered: marginal maximum likelihood (MML) and conditional maximum likelihood (CML). For incomplete designs, the study considered three designs: random incomplete, MST, and targeted testing designs. For each of the six resulting conditions, the study provided an example and discussed results in terms of SE of item parameter estimates under the Rasch model.

Key Findings: The study found that the MML method could be applied across all three incomplete designs, though it occasionally encountered issues with targeted testing designs. However, the performance of the CML method depended on the type of incomplete design: it was feasible for random incomplete and targeted testing designs, but generally not appropriate for MST designs, particularly with small sample sizes. Additionally, the study emphasized that, when a sample is assumed to be random, this assumption should be verified before using the MML method, as violations could cause the algorithm to fail in producing valid results.

Key Limitations: The study focused exclusively on dichotomously scored items using the Rasch model. Future research could extend this work by considering more complex IRT models and incorporating other types of items.

Gierl, M. J., Lai, H., Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2-3), 188-203.

Description: The study evaluated the performance of Computer Adaptive Testing - Simultaneous Item Bias Test (CATSIB) for detecting differential item functioning (DIF) within an MST framework. The study considered a 7-stage MST design with one module at the first stage and three modules (easy, medium, and hard) in subsequent stages. Each module consisted of four items, resulting in a total of 28 items administered per examinee. For simulation, the study considered three sample size conditions: 1,500, 3,000, and 4,500 examinees per reference and focal groups. These conditions corresponded to 100-175, 200-300, and 300-450 examinees per group responding to each item. Moreover, the study implemented both balanced and unbalanced designs. Balanced designs had equal sample size ranges for reference and focal groups, where unbalanced designs had unequal ranges. Routing from stage to stage was based on NC scores: examinees scoring 0–1, 2–3, and 4 were routed to the easy, medium, and hard modules, respectively. CATSIB performance was evaluated in terms of type I error rate and power at a nominal level of 5%.
Key Findings: CATSIB performed well, maintaining appropriate type I error rates and demonstrating adequate power for moderate-to-large and large-to-large sample sizes in the reference and focal groups. The study suggested that, within the MST contexts, CATSIB could be used for detecting DIF items when sample sizes are moderate or large.
Key Limitations: Future research could employ conditional SEM to enhance the accuracy of ability estimates. Additionally, the study highlighted the importance of investigating the underlying reasons for DIF occurrences detected by CATSIB.

Glas, C. (1988). The rasch model and multistage testing. *Journal of Educational Statistics*, *13*(1), 45-52.

Description: This study examined and compared two methods for estimating item parameters within the framework of MST: the CML and MML methods, both presented in the context of the Rasch model. Since the CML method cannot calibrate all items on a single latent continuum, it requires separate calibration for each subgroup. In contrast, the MML method assumes that subjects represent a population with a specified ability and calibrates all items simultaneously.

Key Findings: Due to the assumption about the ability distribution in the MML method, the CML method is generally more powerful in incomplete designs. However, in MST, the MML method could produce solvable equations, whereas the CML method failed to generate solvable equations that could place item parameter estimates on a common scale.

Key Limitations: The author does not explicitly state limitations of the study or research methodology.

Hambleton, R. K., Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education*, *19*(3), 221-239.

Description: This study investigated the impact of optimal and nonoptimal CBT designs on pass-fail decision-making in the context of a credentialing examination. Three CBT designs were considered: linear parallel-form test (LPFT), MST, and CAT. The item bank consisted of 600 dichotomously-scored operational items that were calibrated using the 3PL IRT model. In order to make pass-fail decisions, the study considered three passing scores of θ = -0.5, 0.0, and 0.5. For LPFT, the study constructed five non-overlapping parallel forms of 60 items, with target TIFs centered at each passing score. For MST, a 1-3-3 panel was assembled with 20 items per module; TIFs for the 60 items were centered at 0.0 or 0.5. Five parallel forms were assembled for the routing module, and two parallel forms for each of the six modules at the second and third stages. For CAT, test lengths were also fixed to 60 items, with items selected to maximize information at provisional abilities while satisfying content and conditional/overall item exposure constraints. Test designs were considered optimal when the center of the TIF matched the passing score or the mean of the ability distribution; otherwise, they were considered nonoptimal. For each CBT design, 5,000 simulees were generated from $N(0, 1)$, abilities were estimated using the EAP method. Results were evaluated in terms of decision accuracy (DA) and decision consistency (DC).

Key Findings: The three test designs outperformed the baseline results obtained from randomly-selected 60 items. Overall, CAT produced the best outcomes among the designs. For LPFT, performance improved when TIF was centered on the mean of the ability distribution rather than on the passing scores. When the center of a TIF was aligned with the passing score, MST slightly outperformed LPFT. However, differences in DC and DA among test designs were not substantial.

Key Limitations: The simulations were not extreme or varied enough to fully capture the potential benefits of individualized designs like CAT and MST on DA and DC. The focus on pass-fail decisions may have understated CAT's advantages, as its strengths in proficiency estimation do not always translate to a single decision. Additionally, the simulations may not fully reflect practical challenges in constructing multiple LPFT exams optimally around passing scores.

Han, K. T. (2020). Framework for developing multistage testing with intersectional routing for short-length tests. *Applied Psychological Measurement*, *44*(2), 87-102.

Description: The study proposed a new approach for MST with intersectional routing (ISR). In this approach, final score estimates from a previous section are used to generates initial estimates for a subsequent section measuring a different trait, allowing examinees to be assigned to more appropriate modules at the first stage. The study conducted two simulation studies measuring two traits and employed 2-3 and 3-4 panel structures for the second trait. For 100,000 examinees, the two latent trait scores were generated from a bivariate normal distribution with each trait following $N(0, 1)$. For correlation between the two traits, three levels were considered: .3, .5, and .7. The study used a 14-item fixed form to estimate the first trait, and a simple regression model provided initial estimates for the second trait. Baseline conditions included two scenarios: (1) a typical MST design with a routing module at the first stage, and (2) MST with ISR using true latent scores for routing for the second section. Performances of MST with ISR were evaluated in terms of reliability and the accuracy of ability estimates (MAE and bias). Furthermore, the study explored polynomial and multiple regression models for generating initial estimates for the second trait.

Key Findings: For both panel structures, the MST with ISR designs yielded lower MAEs than the typical MST designs. Within the MST with ISR framework, both estimation accuracy and measurement efficiency improved as the correlation between the two traits increased; and the MAEs for the 3-4 design were also smaller than those for the 2-3 design. When obtaining ISR scores, multiple regression models led to a higher percentage of optimal paths with lower MAEs, whereas the polynomial regression did not show this improvement.

Key Limitations: For MST with ISR, prediction error could increase the possibility of suboptimal routing compared to typical MST designs. Therefore, the potential negative effects of suboptimal routing should be considered and mitigated in operational implementations. Additionally, since the study examined a limited set of factors and conditions, the generalizability of the findings should also be limited.

Han, K. T., Dimitrov, D. M., Al-Mashary, F. (2019). Developing multistage tests using D-scoring method. *Educational and Psychological Measurement*, *79*(5), 988-1008.

Description: The study developed two approaches for implementing the D-scoring method in MST and proposed two criteria for selecting subsequent modules. For selecting subsequent modules, the study considered two criteria: (a) $MinD\delta$ with minimized distance between interim D-score estimate and the average of expected item difficulties and (b) $MinDb$ with minimized distance between D-score estimate and the averaged item location from a logistic regression model. In estimating D-scores, the denominator was determined using two different MST paths: (1) the unique path and (2) the hardest path. For simulation, the study employed 1-3 and 1-2-3 designs, representing short (30-item) and long (60-item) tests with 10 and 20 items per stage, respectively. Baseline conditions included two linear test forms that consisted of items from the routing-easy and routing-easy-medium paths for the 1-3 and 1-2-3 designs, respectively. For each study condition, D-scores were generated for 10,000 subjects based on true D-scores drawn from $\theta \sim N(0, 1.5)$. The performance of the D-scoring method in MST was evaluated in terms of the accuracy of D-score estimates (conditional/overall bias and MAE) and rank order preservation (correlation between true and estimated scores).
Key Findings: Overall, for the two different MST designs, the D-score recovery were better than those for the linear tests. The study also found that using the unique path helpted control bias in D-score estimates more effectively, ensuring score comparability across different paths. For selecting subsequent modules, the $MinD\delta$ criterion slightly outperformed the $MinDb$ criterion.
Key Limitations: In MST with the D-scoring method, it is crucial that items in the routing stage adequately represent the overall test to ensure score comparability across different paths. Future research should explore how the D-scoring method could be applied to item-level CAT where test forms may differ completely for each test taker.

Han, K. T., Guo, F. (2013). *An approach to assembling optimal multistage testing modules on the fly* (GMAC Research Report No. RR-13-01). Reston, VA: Graduate Management Admission Council.

Description: The study introduced a new MST by shaping (MST-S) method and compared its performance to those for CAT and traditional MST (MST-R). For each examinee, the new MST-S method assembles "on the fly" modules, aiming to fill the gap between the current and target TIFs at an interim ability estimate for subsequent stages. For simulation, the study constructed three parallel 1-3-3 MST-R panels, with 20 items per module. Examinees were routed to the module providing the most information at their interim ability. For CAT, two methods were considered for selecting subsequent items: (1) MFI in conjunction with the randomesque method, and (2) a-stratification with b-blocking. For MST-S, the study considered a three-stage design with 20 items per stage, and three iteration levels (3, 6, and 100) were tested to reach the targeted information. For each study condition, 60,000 simulees were generated from a uniform distribution ranging from -3 to 3. Interim and final abilities were estimated using the EAP and MLE methods, respectively. Results were evaluated using conditional error statistics, including SE, MAE, and bias.

Key Findings: The study demonstrarted the feasibility of the MST-S approach. After three iterations of module-shaping process, the measurement precision of MST-S was comparable to that of MST-R. Conditional error statistics for MST-S were higher than those for CAT; however, it was apparent that increasing the number of iterations could reduce these errors. Additionally, the study showed that MST-S could enhance the item pool utilization while maintaining control over item exposure rates.

Key Limitations: Although MST-S could be efficient for MST assembly, it should not be considered as a "one-size-fits-all" solution. Therefore, Testing programs should first clarify their specific goals before selecting a testing design. Furthermore, it would be worthwhile to explore how MST-S could be integrated with other ATA approach.

Hembry, I. F. (2014). *Operational characteristics of mixed-format multistage tests using the 3PL testlet response theory model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: This dissertation explored the administration of mixed-format testlet-based MSTs using the 3PL TRT model, which accounted for the local dependency among items sharing the same testlet while preserving the unique response patter of each item. Item responses from a large-scale assessment were used to estimate parameters under the 3PL TRT model, which was also used to generate responses for simulation. For simulation, the study considered four MST panel designs (1-3, 1-5, 1-3-3, and 1-5-5), two test lengths (55 and 44 items), three routing procedures (AMI, stage-level DPI, and module-level DPI), and three levels for the local item dependence (0, 0.8, and an estimate corresponding to the full size item pool). Modules were assembled to achieve target TIFs at specified ability points. For each study condition, the study generated 100 data sets, each with 1,000 simulees from $N(0, 1)$, and both interim and final abilities were estimated using the EAP method. Results were evaluated in terms of the accuracy of ability estimates (bias, RMSE, MAD, correlation between true and estimated abilities) and module exposure rates.

Key Findings: Based on the study results, bias was minimal across all testing conditions. Among the MST designs, measurement precision tended to improve with fewer stages, longer test lengths, and lower levels of local item dependence. However, the routing procedure had little effect on measurement precision.

Key Limitations: For a fixed test length, the two-stage MST designs included more items per stage than the three-stage MST designs, potentially providing more information at each stage and resulting in higher classification accuracy for the two-stage MSTs. The study considered only a limited range of test lengths; future research could explore shorter tests to determine the minimum length required for mixed-format testlet-based MST. Additionally, all simulees were generated from $N(0, 1)$, which may not capture the full range of realistic ability distributions.

Hendrickson, A. (2002). *Scaling of two-stage adaptive test configurations for achievement testing* (Unpublished doctoral dissertation). The University of Iowa.

Description: The study examined whether a two-stage testlet-based test could balance the advantages of adaptive testing–more precise measurement–with those of conventional linear tests–greater control over test characteristics. The study utilized operational data from Form K of the Iowa Tests of Basic Skills (ITBS) battery, which included four tests: Vocabulary, Reading, Language, and Math. The study used both the 3PL and graded response models for calibration and scaling. For each content and scaling model combination, the study constructed a scaling test spanning grades 3 through 8 and developed a raw-to-scale score conversion table. Additionally, for each combination, all possible configurations of 1-2 and 1-4 MST designs were constructed based on seven groups of items, each group covering two adjacent grades when possible. Test scores were converted to scale scores by using true score equating and two alternative scaling procedures. For each configuration, results were evaluated in terms of psychometric properties (bias, conditional SEM, RMSE, and information), cutscores, and content/process skill analysis.
Key Findings: For all four content areas, the two-stage testing design could improve the psychometric properties of test scores. However, the results seemed to be affected by the choice of scaling model and the evaluation criterion.
Key Limitations: Since examinees did not actually take the test, the results should be considered only as approximations. The study also did not address practical issues such as item presentation on computer screen, selection of the first-stage test, routing examinees with scores near cutscores, time limits, cheating, or accommodating examinees with learning disabilities. Additionally, the items and item groups used to assemble the two-stage tests were drawn from linear tests and were not originally designed for adaptive testing. Furthermore, cutscores were based on estimated scores from the first-stage test; simulation studies could provide a more rigorous basis for determining cutscores.

Hendrickson, A. (2007). An ncme instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

Description: This article described two-stage, testlet-based MST and examined its relative advantages and disadvantages compared to P&P and item-level adaptive tests. It also discussed key issues and outlined the steps involved in developing MST.

Key Findings: Compared to P&P tests, MST offers several advantages: 1) more efficient and precise measurement; 2) reduced testing and score reporting time while maintaining equal or higher predictive and concurrent validity of score inferences; and, 3) greater flexibility in scheduling testing. Compared to item-level adaptive tests, MST provides advantages in the five aspects: 1) better assurance of unidimensionality of the test and local independence between testlets; 2) use of item sets and greater control over item ordering, which may reduce context effect; 3) increased control over nonstatistical test properties, enhancing overall test quality; 4) improved test security through controlled item and test exposure; 5) the ability for examinees to review items; and, 6) accelerated scoring with reduced demands on routing, data management, and computer processing. The article also discussed both limitations and disadvantages of MST. MST requires more items to achieve the same measurement precision and more effort from item writers and test editors. Also, it would be more challenging to replace an entire item set, and routing errors could be larger than those in item-level adaptive tests due to fewer adaptation points. Furthermore, the author highlighted the critical considerations in constructing a MST program, including the number of stages and testlets, targeted statistical and qualitative specifications, and methods for scoring, adapting, and assembling modules and panels.

Key Limitations: This article focuses on describing two-stage and testlet-based MST and does not explicitly state limitations of the study or research methodology.

Jewsbury, P. A., van Rijin, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics*, *45*(4), 383-402.

Description: When tests are designed to measure multiple dimensions, with each item assessing a single dimension, item parameters can be estimated either using a single multidimensional IRT (MIRT) model or using multiple unidimensional IRT (M-UIRT) model (one UIRT model per dimension). This study investigated this setting within the framework of missing data mechanisms–MCAR, MAR, and MNAR–and demonstrated that M-UIRT models could produce strongly biased item parameter estimates both mathematically and empirically. For simulation, the study used items from the 2017 NAEP Grade 4 Mathematics nonadaptive computer-based operational assessment which measured 5 dimensions, and constructed a 1-2 MST design. For each simulee, five abilities were generated from a multivariate normal distribution using the latent covariance matrix derived from the real data. The study generated 100 data sets, each with 10,000 simulees. The NC cutscore of 13 was applied to route simulees to either the harder or the medium-difficulty block in the second stage. Results were evaluated based on mean difference (MD) and RMSE in item parameter estimates. Furthermore, using real data, item parameter estimates obtained with M-UIRT models were compared to those obtained using a MIRT model.

Key Findings: The use of M-UIRT model resulted in biased item parameter estimates as demonstrated both mathematically and empirically. Furthermore, the bias exhibited a consistent direction and magnitude across items within the same block, indicating that ability estimates could also be subject to bias. Therefore, although using M-UIRT models could be considered acceptable for nonadaptive tests, their application in adaptive MST contexts lacks sufficient justification.

Key Limitations: The study focused on two-stage MST, and future research should consider more complex designs or situations where items are scored differently for the routing decisions and for final item calibration (e.g. automatic vs. human). Furthermore, reference item parameters were derived from a nonadaptive test and a different sample, suggesting that differences between MST estimates and reference values could be due to sample and contextual differences rather than M-UIRT bias alone. This issue also led to the exclusion of outlier items exhibiting extreme differences, likely due to these contextual effects.

Jiang, Y. (2019). Statistical considerations for subscore reportingin multistage testing. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 129-136). Springer Nature Switzerland AG: Springer.

Description: The study investigated factors that could impact the accuracy of subscore estimates and subscore reliability in the context of MST. The study first generated multiple item pools by varying the percentage of items required for subtest assembly (70% and 50% usage), average item discrimination (.7, .8, .9, and 1.0), and the distribution of item difficulty parameters ($N(9, .7^2)$, $N(0, 1)$, and $N(0, 1.3^2)$). For each item pool, the study conducted simulation by varying the number of subtests (2 and 3), subtest length (16 and 20 score points), and correlation among subscores (.5, .7, and .9). For each study condition, the study generated 100 data sets, each with 4,000 examinees from a multivariate normal distribution. A 1-3 MST panel was assembled based on ATA using *lpSolve* software. Both interim and final ability estimates were obtained using the EAP method. From the first to second stages, cutscores for routing were determined so that approximately 22% to 25% of examinees were routed to the easy and hard modules; the rest were routed to the medium-difficulty module. Results were evaluated in terms of the accuracy of subscore estimates and their reliability.

Key Findings: The study results showed that the subscore reliability and estimation accuracy improved with longer subtests, larger and more discriminating item pools, smaller variability in item difficulty, and lower correlation among subscores.

Key Limitations: The study was limited to dichotomously-scored items and subtests of simple structures. Future research should incorporate polytomously-scored items into item pools and examine subtests with complex structures. Moreover, since subscores should have added value, future research should also identify the conditions under which reporting subscores in MST could provide additional information about examinees' abilities.

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pools* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Description: This study compared the performance of three test designs: LFT, MST, and CAT. For simulation, the study varied conditions for four study factors: 1) item pool quality (low, moderate, and high discriminating), 2) alignment between test and item pool content specifications (proportional and disproportional), 3) test length (40 and 60 items), 4) exposure control (three levels for each design). For MST, one additional factor, panel design, was considered with two configurations: 1-3 and 1-3-3 designs, each with an equal number of items per module. For each of the six combinations regarding the item pool characteristics, a pool of 450 items was generated using the 3PL IRT model. For each crossed study condition, 9,000 examinees were randomly simulated from $N(0, 1)$. For MST, routing was completed based on cutscores of -.43 and .43; for CAT, items were adaptively selected to maximize information while meeting content and exposure constraints. Across all designs, abilities were estimated using the EAP method. The designs were evaluated in terms of overall and conditional measurement precision as well as classification accuracy at cutscores corresponding to 30%, 50%, and 85% failure rates.

Key Findings: Across all designs, test reliability, overall and conditional measurement precision, and classification accuracy improved with higher-quality item pools (i.e., higher discriminating), longer test length, and better alignment between the item pool and test specification (i.e., proportional matching). Performance declined as the maximum exposure rate decreased and as exposure control became more restrictive. Overall, with the other conditions being fixed, CAT demonstrated better psychometric properties compared to MST, which in turn outperformed LFT.

Key Limitations: When item pool utilization was high, the test assembly software exhibited limited capacity to fully satisfy the specified test requirements.

Jodoin, M. G., Zenisky, A., Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203-220.

Description: This study compared the psychometric properties of several CBT designs in the context of credentialing examinations, where pass-fail decisions demand high levels of decision accuracy and measurement precision. Performances for four operational 60-item tests were compared to: (a) three new 60-item LFT forms, (b) six 60-item 1-3-3 MST forms, and (c) six 40-item 1-3 MST forms. The study considered 3 passing scores that produced pass rates of approximately 30%, 40%, and 50%. For the 1-3-3 MST forms, the study considered two conditions for the target information function (TIF) designs: (1) each module in Stages 1, 2, and 3 was assigned to one-third of the mean TIF from the operations forms, and (2) Stage 1 modules and Stage 2 and 3 modules were assigned to one-quarter and three-eighths of the mean operational TIF, respectively. For each of the TIF designs, the study assembled three parallel modules in Stage 1 and constructed three panels to maintain an item exposure level of 33%. The 1-3 MST forms were constructed by removing Stage 3 modules from the 1-3-3 MST designs. For simulation, the study used item parameters from a large-sample based operational forms that were calibrated under the 3PL IRT model. For each test form, 5,000 simulees were randomly generated from $N(0, 1)$. Abilities for scoring and routing were estimated using the MLE method. The designs were evaluated in terms of the accuracy of ability estimates, classification accuracy, and decision consistency.

Key Findings: The study results showed that, regardless of design or passing score, all 60-item tests produced more accurate ability estimates, with decision consistency and decision accuracy demonstrating similarly acceptable levels. The 40-item test forms also produced results within an acceptable range. Additionally, the performances of the two TIF designs were also very comparable.

Key Limitations: Since the item bank was constructed from the operational forms, it was too realistic to meet all content and statistical specifications for the exams under study. Future research could employ a simulated item bank, which would allow for more ideal content and statistical specifications and might be able to provide advantages of MST designs more precisely.

Kaplan, M. (2016). *New item selection and test administration procedures for cognitive diagnosis computerized adaptive testing* (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey.

Description: This dissertation conducted three separate studies, each addressing a specific objective. The first study introduced two new methods for selecting items in cognitive diagnosis CAT (CD-CAT): (1) the modified posterior-weighted Kullback-Leibler (MPWKL) index, and (2) the G-DINA model discrimination index (GDI). Within the context of CD-CAT testing, the study conducted simulations incorporating various factors, including item quality, generating model, and test termination rule. To evaluate the efficiency of the proposed methods, the study compared the correct classification rates for attribute and attribute vector, administration time, and item usage of the two new methods against those obtained using the posterior-weighted Kullback-Leibler (PWKL) index.

The second study examined the efficiency of the GDI method under two item exposure control strategies: (1) the restrictive progressive (RP) method and (2) the restrictive threshold (RT) method. A simulation study was conducted considering various factors, including item quality, generating model, attribute distribution, item pool size, sample size, and pre-specified desired exposure rate. To assess the efficiency of the GDI, results were evaluated in terms of item exposure rates and the estimation accuracy.

The third study introduced a blocked-design procedure for CD-CAT testing, where examinees could review items and change responses during test administration. The study proposed four versions of the blocked-design procedure: unconstrained, constrained, hybrid-1, and hybrid-2. The study conducted simulation by varying conditions for item quality, generating model, block size, and test length. Items were selected using both the GDI and the PWKL index, and the efficiency of these indices were evaluated based on the correct attribute classification (CAC) rate and the correct attribute vector classification (CVC) rate.

Key Findings: Based on the results of the first study, the performances of the MPWKL index and GDI were highly comparable. Compared to the performances of the PWKL index, both methods demonstrated higher correct attribute classification rates and/or shorter mean test length. Additionally, among the three indices, the GDI required the least time for test administration. The second study showed that the GDI performed efficiently under both item exposure control methods, with the RP method producing more uniform item exposure rates than the RT method. Moreover, when the RT method was used, fewer factors had substantial impact on exposure rates compared to the RP method. Overall, the findings suggested that using the RP method in conjunction with the GDI seemed more promising for operational uses. Based on the results of the third study, the new procedures in conjunction with the GDI performed reasonably

well although there was a slight reduction in CAC and CVC rates.

Key Limitations: All three studies employed a Q-matrix that included all possible q-vectors, which does not reflect realistic settings in operational testing. Additionally, the studies examined a limited number of study factors and conditions. Future research should consider a broader range of conditions to enhance the generalizability of the findings to other settings.

Karatoprak Ersen, R., Lee, W. (2023). Pretest item calibration in computerized multistage adaptive testing. *Journal of Educational Measurement*, *60*(3), 379-401.

Description: The study compared the performance of various calibration and linking methods for pretest items in MST. For placing pretest items, the study considered two models: (1) the embedded-section (ES) model administering pretest items in a separate module and (2) the embedded-item (EI) model distributing pretest items across modules. For simulation, the study used a 40-item 1-3 design, with module lengths of 20-20 and 10-30. Also, the study varied conditions for the number of pretest items (12 and 20), sample sizes (1500 and 3000), and ability distribution ($N(0, 1)$, $N(.5, 1)$, $N(-.5, 1)$, and $N(.5, .64)$). Interim abilities were estimated using the EAP method and cutscores for routing were determined by the AMI method. Using the 3PL model, the study implemented two calibration methods: (1) separate calibration (SC) and (2) fixed calibration (FC). To place pretest item parameter estimates on the scale of an item pool, the study considered three different sets of items for linking: (1) operational items in the routing module (R-items), (2) R-items but with the remaining operational items freely estimated, and (3) operational items from all modules. Performance was evaluated with respect to item parameter recovery and the ICC for pretest items.

Key Findings: The performances of the SC method consistently outperformed and exhibited greater stability than the FC method. Linking pretest items using all operational items from all modules produced the most accurate parameter estimates for pretest items. Additionally, the ES model yielded more accurate and stable results compared to the EI model.

Key Limitations: The study was limited in terms of factors and conditions, which constrains the generalizability of the findings. This study considered one-time linking for pretest items. Since those pretest items could be reused for linking subsequent pretest items, future research should examine error accumulation and potential scale drift to more comprehensively evaluate different calibration and linking approaches. Additionally, future studies could explore alternative IRT models, other item types, and item pools with content specifications.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: This study compared the performance of item-level CAT (adaptive at testlet and item levels), testlet-level CAT (adaptive at testlet level), and 1-3-3 MST with respect to measurement effectiveness and exposure control properties. A simulation study was conducted by manipulating item pool size (full and reduced), ability distribution (normal and negatively-skewed), and test length (21 and 42 items). The full-sized item pool consisted of 1,008 items drawn from an entire set of testlets and items from a large-scale assessment, while the reduced-sized pool contained 672 items, representing two-thirds of the full pool. Responses were generated using the 3PL TRT model; and, abilities were estimated using the EAP method. The performance of the three adaptive test designs was evaluated in terms of the accuracy of ability estimates (bias, RMSE, and MAD) and exposure control properties (item and testlet exposure rates).

Key Findings: Measurement accuracy was comparable and satisfactory across all three test designs. Testlet-level exposure properties were also acceptable for the three designs. However, when the exposure control properties were examined at the item level, the results varied by test designs. Testlet-level CAT exhibited the most favorable item-level exposure performance. For item-level CAT, pool utilization was suboptimal, although item exposure rates remained within the pre-specified maximum limit. In contrast, MST demonstrated excellent pool utilization, but with a relatively high proportion of items with high exposure rates. Additionally, exposure control properties were noticeably deteriorated under the skewed ability distribution.

Key Limitations: The study employed the 3PL TRT model only; future research could consider polytomous IRT models for testlet-based items. Although results were similar across the full and reduced item pools, the reduced item pool might not have been sufficiently small to produce meaningful differences. Furthermore, the study examined one specific MST design, suggesting that the results should not be generalized to other MST configurations.

Kim, H., Plake, B. S. (1993). *Monte carlo simulation comparison of two-stage testing and computerized adaptive testing.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Description: The study compared CAT and two-staged MST with respect to measurement accuracy and efficiency. For the two-staged MST designs, the study varied conditions for three factors: the distribution of item difficulty at the routing stage (peaked and rectangular), the length of the routing test (10, 15, and 20 items), and the number of 30-item tests at the second stage (6, 7, and 8). For simulation, the study generated separate item pools for CAT and MST by varying the distributions of difficulty parameters. In MST, the difficulty distribution depended on the characteristics of the routing tests, and examinees were routed subsequent tests whose average difficulty was closest to their interim ability estimates. For CAT, a uniform distribution from -3 to 3 was used to create an item pool, and subsequent items were selected to provide maximum information, with testing terminating after 40, 45, or 50 items. Throughout the study, the modified 1PL IRT model was implemented and abilities were estimated using the MLE method. For each study condition, 100 simulees were generated at each of 16 quadrature points ranging from -3 to 3 in increments of .4. Performances of CAT and MST were evaluated in terms of correlation coefficients and the accuracy of ability estimates (bias and RMSE).

Key Findings: Regarding measurement accuracy and efficiency, fixed-length CAT outperformed the two-stage MST designs of equivalent length. Within the MST designs, routing tests with a uniform difficulty distribution produced more accurate ability estimates than those with a peaked distribution. Furthermore, the two-stage tests with an odd number of second-stage tests demonstrated superior performance compared to other two-stage configurations.

Key Limitations: Since the results could depend on specific study characteristics (e.g., item pool and IRT model), the generalizability of the findings should be limited within the scope of this study.

Kim, J. (2010). *A comparison of computer-based classification testing approaches using mixed-format tests with the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: This study employed the GPCM for mixed-format tests and compared the performance of CAT, MST, and sequential probability ratio test (SPRT). For simulation, the study varied conditions for test length (21, 27, and 33 test units) and cutscores (-0.524, 0.000, and 0.524). To control exposure rate, both the CAT and SPRT approaches used a progressive-restricted exposure control procedure with a pre-specified maximum test unit exposure rate. For each study condition, the study generated 100 data sets, each with 1,000 simulees drawn from $N(0, 1)$. For MST, easy, medium, and difficult modules were assembled to achieve target TIFs at θ = -1, 0, and 1, respectively. Abilities were estimated using the MLE method, and examinees were routed to subsequent modules providing maximum information at their interim ability estimates. For SPRT, nominal type I and type II error rates were set at 0.05, with an indifference region width of 0.50 ($\theta \in [-.25, .25]$). Performance was evaluated in terms of classification accuracy and item exposure control.

Key Findings: All three methods demonstrated strong performance in classifying examinees into two categories. The performances for the CAT and SPRT designs were comparable and generally superior to that of the MST design. For all designs, the classification accuracy improved as test length increased. However, predicted classification decision seemed to depend on the location of the cutscores. Regarding exposure control properties, the CAT method achieved lower test unit exposure rate and more efficient pool utilization rate compared to the SPRT method, while the MST design exhibited the worst performance.

Key Limitations: The study considered only two categories for classifying examinees. Additionally, the item pool was unevenly distributed with respect to the maximum information across ability points, which in turn posed challenges in constructing easy modules. Finally, the use of a single MST design restricted the generalizability of the findings to other MST configurations.

Kim, J., Chung, H., Dodd, B. G., Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, *72*(4), 574-588.

Description: The study compared various panel designs of mixed-format MST for pass/fail classification tests. For MST, a simulation study examined three factors: (1) TIF at the first stage (three levels), (2) the center of the first-stage TIFs (three levels), and (3) passing rate (three levels). The test unit pool for the study included 424 test units across nine content cells, three test unit types (dichotomous, three categorical, and four categorical scores) and three subcontent areas (content I, II, and III). All item parameters were calibrated using the GPCM. The study constructed a 1–3–3 panel structure, with nine test units per module, yielding 43 possible points for total scores. Routing was completed using the modified AMI method. For baseline comparison, CAT simulation was conducted by manipulating conditions for two factors: (a) exposure control (two levels) and (b) passing rate (three levels). In CAT, subsequent test units were selected based on the MI or randomesque-10 methods. Across all conditions, abilities were estimated using the MLE method. For each condition, the study generated 40 data sets, each with 1,000 simulees randomly generated from $N(0, 1)$. Performances of the MST and CAT conditions were evaluated in terms of correct classification rate (CCR), false-negative error rate, false-positive error rate, and total error rate.

Key Findings: The study found that all MST conditions performed well. CCRs were higher when higher levels of TIFs were assigned to the first-stage modules. Under these conditions, MST results for were comparable to those of CAT with the randomesque-10 procedures. Moreover, MST achieved the best pool usage rates and provided better test security than CAT with the AMI method.

Key Limitations: The pool used in this study did not contain enough easy test units (i.e., the test unit pool was somewhat negatively skewed), which made assembling easy modules challenging. Future research could explore test unit pools with different distributions of item difficulty and examine how these characteristics interact with MST panel constructions.

Kim, J., Chung, H., Park, R., Dodd, B. G. (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.

Description: The study compared MST designs using different routing methods for classification testing. For simulation, the study varied panel structure, routing method, test length, and passing rate. The item bank consisted of 157 items calibrated under the PCM; each item had three, four, or five score categories. For panel structure, the study considered 1-2-2, 1-2-3, 1-3-2, and 1-3-3 designs. Each module included either three or five items, yielding test lengths of nine items (23 score points) or 15 items (38 score points). For each structure, the study assembled three parallel panels and randomly assigned one panel to examinees. For routing, the stage-level DPI (SL-DPI), the module-level DPI (ML-DPI), and the modified AMI (M-AMI) methods were considered. Abilities were estimated using the MLE method. Pass/fail decisions were then made using three targeted passing rates (20%, 50%, and 80%). The corresponding cutscores were determined based on the standard normal distribution, $N(0, 1)$. For each condition, the study generated 50 replications, each with 1,000 examinees randomly generated from $N(0, 1)$. Results were evaluated in terms of decision precision (classification accuracy and error rates) and test security level (item exposure rate and pool utilization).

Key Findings: For a given test length and panel structure, all routing methods performed similarly. All panel structures also showed comparable performance when test length and routing method were held constant. However, the number of modules at the second or third stages did not matter much. The study also showed that the panel structures with a longer test length produced better results than shorter ones. Regarding item exposure rate, the two DPI methods performed similarly, while the M-AMI method gave slightly higher variability in item exposure rate. Pool utilization was excellent across all MST conditions.

Key Limitations: The findings should be interpreted within the specific scope of this study. Future research could consider other conditions, such as larger item pools, two-stage panels, and modules with reduced overlap.

Kim, S., Livingston, S. A. (2017). *Accuracy of a classical test theory–based procedure for estimating the reliability of a multistage test* (ETS Research Report No. RR–17-02). Princeton, NJ: Educational Testing Service.

Description: This study examined the accuracy of a CTT-based procedure for estimating the reliability of scores on a multistage test. For simulation, the study used a 1-2-3 panel design and generated 10 parallel forms using items simulated under the 2PL model. Each module contained 15 items. Routing was completed based on NC cutscores; for each form, cutscores were determined so that approximately equal proportions of examinees could follow each possible path. Furthermore, for each of the 10 forms, the study conducted IRT true-score equating and constructed four conversion tables to convert NC scores to scale scores; there was one conversion table associated with each path. A group of 30,000 examinees were randomly generated from $N(0, 1)$ and completed all 10 forms. For each form, the study computed both CTT-based and IRT-based reliability estimates in scale scores, as well as all possible correlations of scale score estimates across forms.

Key Findings: The CTT-based method produced highly accurate results, consistent with findings from a previous study on two-stage MST (Livingston, & Kim, 2014). These results provide further evidence supporting the generalizability of the CTT estimation approach. Compared to the IRT-based method estimation method, reliability estimates from the CTT-based estimation method were slightly less accurate .

Key Limitations: Since responses were generated using the IRT model, it is unsurprising that the IRT-based method gave higher reliability estimates than the CTT-based method.

Kim, S., Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (ETS Research Report No. RR–14-01). Princeton, NJ: Educational Testing Service.

Description: This study examined the impact of misrouting in a 1-3 MST design. For routing, the study considered two scenarios: (1) routing examinees to a single target module at the second stage, and (2) assigning all examinees to all three second-stage modules and obtaining equated raw scores for each paths. To route examinees to a target module, two cutscores were determined using the DPI method so that approximately 30%, 40%, and 30% of examinees were routed to difficult, medium, and easy modules, respectively. The study also included two conditions for MST form assembly: (a) a small-difference condition where second-stage modules overlapped in difficulty, and (b) a large-difference condition where modules differed substantially in difficulty. One thousand examinees were simulated at each of 41 quadrature points ranging from -3 to 3 in increments of .15, and completed the two MST forms. For each form, differences in examinees' equated scores across the three different paths were evaluated in terms of conditional mean difference, SE of the mean difference, and z-statistics at each of the quadrature points. Additionally, the study subtracted equated-scores for the target modules from equated-scores for the one-level off modules, and these difference scores were evaluated in terms of conditional mean difference and RMSE at each of possible NC score points for the routing module.

Key Findings: Regardless of the degrees of overlap in difficulty among the second-stage modules, score differences (i.e., bias) across paths were negligible for practical purposes. Overall, the impact of misrouting was minimal under the MST designs examined in this study.

Key Limitations: Future research could explore other factors, such as measurement conditions at the routing stage (e.g., optimal number of items per routing module), scoring methods, and differences in item discrimination among second-stage modules, to investigate their effects.

Kim, S., Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (ETS Research Report No. RR-16-22). Princeton, NJ: Educational Testing Service.

Description: This study evaluated the robustness of IRT proficiency estimation methods to atypical responses. Using the revised GRE design, the study assembled a 1-3 MST panel based on the 2PL IRT model. For routing, the study selected two cutscores using the DPI method so that 30%, 40%, and 30% of the examinees could be routed to the difficult, medium, and easy modules at the second stage, respectively. For each of seven atypical response types considered in the study, 1,000 examinees were simulated at each of the 41 quadrature points ranging from -3 to 3 in increments of .15. Both interim and final abilities were estimated using one of five methods: (1) TCC with NC scoring (TCC), (2) MLE with item-pattern (IP) scoring (MLE), (3) EAP with NC scoring (sEAP), (4) EAP with IP scoring (EAP), and (5) maximum (mode) a posteriori with IP scoring (MAP). The performance of these methods was evaluated in terms of the accuracy of ability estimates. Additionally, ability estimates were converted to scale scores, and accuracy in scale scores was compared across the methods.
Key Findings: Regarding the accuracy of ability estimates, the performance of the five methods at the extremes of the theta scale depended on the type of atypical responses. For the random, preknowledge, and peculiar subgroup types, results were similar to those for the no atypical-response condition, with Bayesian methods (EAP, sEAP, and MAP) generally producing smaller overall errors than non-Bayesian methods. For the guessing, careless, creative, and speededness types, bias and RMSEs were larger than those for the no atypical-response condition, and the advantage of Bayesian methods was also minimal. When ability estimates were converted to on scale scores, the performances of the Bayesian methods again performed slightly better than non-Bayesian methods. However, the benefit of using the Bayesian methods were not substantial, leading to the conclusion that the robustness of all five estimation methods to atypical response types was comparable under the conditions considered in this study.
Key Limitations: The study considered atypical response types one at a time. Future research could investigate more complex or problematic atypical responses, explore alternative panel designs, and employ the 3PL model.

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

Description: This study conducted a simulation study to compare the performance of seven IRT proficiency estimators in terms of accuracy of ability estimates. For simulation, the study assembled eight 1-3 panels by varying the difficulty differences between adjacent modules at the second stage (small = .75, large = 1.25) and module lengths (25-15, 20-20, 15-20, and 15-25). The seven proficiency estimators were: (1) TCC with NC scoring, (2) Yen's (1984) MLE with NC scoring, (3) MLE with NC scoring, (4) MLE with item-pattern scoring, (5) EAP with NC scoring, (6) EAP with IP scoring, and (7) MAP with IP scoring. Two thousand examinees were simulated at each of 41 quadrature points from -3 to 3 in increments of .15. For routing, the DPI method was used so that approximately 30%, 40%, and 30% of the examinees could be routed to the easy, medium, and hard modules at Stage 2. Final ability estimates were evaluated in terms of bias, SE, and RMSE.

Key Findings: The choice between Bayesian (EAP and MAP) and non-Bayesian estimators had a greater impact on the accuracy of ability estimates than the choice of NC versus IP scoring. Bayesian methods generally outperformed non-Bayesian methods, particularly for high-performing examinees. Overall, the eight panel designs had minimal effect on ability accuracy. Furthermore, for the large-difference condition, all seven methods performed similarly across the middle of the ability scale from -2 to 2, and measurement precision remained relatively consistent across the entire ability scale. The study also suggested that the impact of the estimator choice is small as long as the test contains sufficient high-quality items matched to examinees' abilities. Additionally, assigning more items to the first stage was found to improve classification accuracy.

Key Limitations: Since scale scores, rather than ability estimates, are typically used for reporting purposes in practice, future research should examine how different estimation methods affect scale scores.

Kim, S., Moses, T., Yoo, H. (2015b). *Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing* (ETS Research Report No. RR–15-11). Princeton, NJ: Educational Testing Service.

Description: The study compared seven IRT proficiency estimators under two-stage 1-3 MST. For two-stage MST designs, the study assembled a total of eight panels varying the difficulty difference between adjacent modules at the second-stage (small and large difference) and module lengths (25-15, 20-20, 15-25, and 10-30). The seven IRT proficiency estimators were: (1) TCC with NC scoring, (2) Yen's (1984) MLE with NC scoring, (3) MLE with NC scoring, (4) MLE with item-pattern (IP) scoring, (5) EAP with NC scoring, (6) EAP with IP scoring, and (7) MAP with IP scoring. Two thousand examinees was simulated at each of 41 quadrature points ranging from -3.0 to 3.0, in increments of .15. For routing, cutscores were determined using the DPI method so that approximately 30%, 40%, and 50% of the examinees were routed to the difficult, medium, and easy modules, respectively. Responses were generated and abilities were estimated using the 3PL IRT model. The performance of the seven estimation methods was evaluated in terms of accuracy of ability estimates (bias, SE, and RMSE).

Key Findings: Overall, the choice between Bayesian (EAP and MAP) and non-Bayesian estimators had a greater impact on ability estimate accuracy than the choice between NC and IP scoring estimators. Bayesian estimators produced slightly lower RMSEs than non-Bayesian estimators. More specifically, all estimators yielded similar RMSEs in the middle of the ability scale; however, for low- and high-performing examinees, different estimators could cause non-negligible score changes. The study also indicated that the impact of estimator choice is minimal when the test contains enough high-quality items matched to examinees' abilities. Additionally, the study found that the accuracy improved with larger difficulty differences among the second-stage modules. And, the percentage of correct classification at Stage 1 increased when more items were assigned to the first stage.

Key Limitations: The generalizability of the findings is limited to the conditions examined in this study. Moreover, in operational testing, scale scores rather than ability estimates are typically reported to examinees. Therefore, it would be worthwhile to investigate the impact of different IRT ability estimators on reported scale scores.

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, *14*.

Description: This article discussed perspectives on CAT from test developers and psychometricians, test-takers, and educators and subject matter experts. The article suggested that the shadow test approach can address CAT limitations related to lack of control over nonstatistical specifications, such as content coverage and balance, item exposure, test length, item format, and item enemies. And, computerized MST was recommended to address CAT's incapability to allow review or modification of answers. Furthermore, the article noted that administering items with a 50% chance of answering correctly in CAT may reduce examinees motivation and leave them feeling discouraged. To investigate this aspect, a simulation study was conducted by manipulating the target probability of answering correctly in the CAT item selection algorithm and varying test length, which produced five configurations.

Key Findings: The findings indicated that, with respect to ability estimation, the performances of the five configurations were almost identical. However, the observed percent correct for each CAT was lower than the target probability of answering correctly, mainly due to a left-skewed distribution of item difficulty in the item bank. The study results suggested that increasing the target probability may help mitigate the negative psychological impact on examinees.

Key Limitations: The author stated that the scope of the paper was limited and pointed readers to external resources for further information.

Li, G., Cai, Y., Gao, X., Wang, D., Tu, D. (2021). Automated test assembly for multistage testing with cognitive diagnosis. *Frontiers in Psychology*, *12:509844*.

Description: The study developed an MST design within the cognitive diagnosis framework (CD-MST). To make the design feasible, the study proposed new indices measuring item difficulty and test reliability. To define item difficulty, the study proposed using the mean correct response probability across all knowledge states (KS) for one item. And, test reliability was quantified using attribute reliability as proposed by Templin and Bradshaw (2013), providing quantitative targets for test assembly. The study also demonstrated how the NWADH algorithm can be applied to assemble CD-MST tests while satisfying both non-statistical and statistical constraints. For simulation, the study employed a three-stage panel, varying test length (21 and 25 items) and the number of parallel panels (5 and 10). To construct panels, the study generated an item pool with 1,000 items with item difficulties computed using the new index. Panels were then assembled to achieve a target reliability of 0.90 while meeting the specified constraints. For five independent attributes, KSs for 1,000 examinees were randomly generated from 32 possible KSs. Results were evaluated in terms of differences between target and observed reliabilities and the number of constraint violations.

Key Findings: The proposed item difficulty index was highly correlated with item difficulty estimates from the Rasch model, supporting its appropriateness as a measure of item difficulty in cognitive diagnosis. Additionally, Cronbach's α coefficients for different pathways exceeded the target reliability of 0.90, suggesting that the new reliability index could ensure reliability and measurement error in CD-MST tests. Furthermore, although some constraints were occasionally violated, the application of the NWADH algorithm successfully produced parallel panels that generally satisfied both non-statistical and statistical constraints.

Key Limitations: Future research could develop alternative indexes for measuring item difficulty in cognitive diagnosis. Also, to establish quantitative targets, future study could also consider other metrics such as classification accuracy, matches, and consistency. Additionally, future studies could explore using different panel assembly methods and varying conditions such as test length.

Linn, R. L., Rock, D. A., Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*(1), 129-146.

Description: The study compared the performances of seven programmed tests, including five routing tests and two branching tests. The five routing tests were constructed based on (1) two-stage, (2) broad range, (3) group-discrimination, (4) four-group sequential item sampling, and (5) three-group sequential item sampling. The two branching tests included: (a) a 10-item test with item branching, where scoring and branching occurred after each item, and (b) a 25-item test with item-blocks branching, where branching occurred after each block of five items. The study used item response data for 190 items answered by 4,885 subjects. Subjects were randomly assigned to an original sample group or a cross-validation sample group, and were administered to the programmed tests. For comparison purposes, five shortened conventional tests (10, 20, 30, 40, and 50 items) were also constructed and administered to the subjects. At the end of testing, total scores on the 190 items were predicted for each programmed test using regression models. Performance was evaluated based on correlations between programmed test scores and (a) the shortened conventional test scores, (b) the 190-item total test scores, and (c) scores on four external tests.

Key Findings: With respect to reproducing total test scores for the 190 items, the programmed tests–particularly the four- and three-group sequential methods and the 5-items-block branching test–slightly outperformed the shortened conventional tests. Additionally, the programmed tests–especially those using group-discrimination and three-group sequential methods, as well as the two branching tests–tended to yield slightly higher correlations with the four external criterion tests compared to the shortened conventional tests.

Key Limitations: The study did not administer the programmed tests operationally; instead, it used responses from the standard P&P format as if subjects had completed the programmed tests. Additionally, items for the programmed tests were not selected using optimal procedures. Future research could consider incorporating ICC-based item selection methods.

Lord, F. M. (1969). *A theoretical study of two stage testing* (ETS Research Bulletin Series No. RB-69-95). Princeton, NJ: Educational Testing Service.

Description: This study examined theoretical aspects of two-stage testing and explored conditions under which two-stage testing could perform as effective as the "best" up-and-down procedures where a single item at a subsequent stage is selected based on responses to previous items (Lord, 1968). For two-stage tests, the study had three conditions: (1) 15-item tests without guessing, (2) 60-item tests without guessing, and (3) 60-item tests with guessing. For simplicity, all items were assumed to have equal discrimination parameters. Each subject was administered to items with equal difficulty parameters. Abilities were estimated using the MLE method. The performance of the procedures was evaluated in terms of the information function as an index of effectiveness in testing. The performance of the two-stage tests was compared to that of a 60-item conventional test and a 60-item up-and-down tailored test.

Key Findings: From theoretical perspectives, the length of a routing test should be neither too short nor too long, and should account for the ability levels that the second-stage tests are intended to differentiate. The number of second stage tests depends on the targeyed ability range and economical limitations. Difficulty levels of the second-stage tests should not be the same or too extreme. For routing, cutscores should be able to assign examinees to second-stage tests with difficulty levels appropriate to their abilities. Also, the study pointed out that equally spaced cutscores generally yielded better results. Based on the simulation studies, two-stage testing can be as effective as the up-and-down tailored testing when items did not involve guessing. However, when there was guessing involved in responding items correctly, none of the two-stage tests matched the best performance of the up-and-down tailored tests, particularly for examinees with near-perfect or near-chance-level abilities. In such cases, measurement accuracy could be improved by increasing the number of second-stage tests or adding additional stages.

Key Limitations: The author was unable to find truly "optimum" designs and instead relied on an iterative process of investigation and modification. Finding optimal cutting points on the routing test proved challenging using trial-and-error methods. Lastly, the current study did not consider shorter tests with guessing. The findings were therefore considered tentative and might not generalize to substantially different contexts.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*(3), 147-151.

Description: To address practical challenges in obtaining comparable scores from *tailored* testing (i.e., two-stage adaptive testing), the author introduced flexilevel tests–self-scoring P&P tests designed to match item difficulty to examinee ability while allowing the use NC scores. In the flexilevel tests, items are arranged by difficulty, with the middle item administered first. Based on the examinee's response, the next item is either a slightly easier or a harder; the test continues until examinees answer half of the items in the whole test. The author also described ten properties of flexilevel tests and provided an example (see pages 150–151 for details).
Key Findings: -
Key Limitations: -

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, *36*(3), 227-242.

Description: The current study aimed to identify effective designs for two-stage testing within specified constraints using a mathematical model grounded in mental test theory. The study explored various two-stage testing procedures (designs), considering a total of nine study factors, including the number of response options per item, the number of alternative second-stage tests available for use, routing test difficulty, cutscores for assigning examinees to second-stage tests (see pg. 228). The study emphasized that "A good procedure should provide reasonably accurate measurement for examinees who would obtain near-perfect or near-zero scores on a conventional test." For two-stage tests, the study considered three conditions: (1) 15-item tests without guessing, (2) 60-item tests without guessing, and (3) 60-item tests with guessing. All items were assumed to have equal discrimination parameters, and items administered at the same stage to a subject had equal difficulty parameters. Abilities were estimated using the MLE method. Test effectiveness was evaluated using the information function. Performance of the two-stage tests was compared with that of a 60-item conventional test and a 60-item up-and-down tailored test.

Key Findings: The study reached two main conclusions. First, when guessing was not involved in answering test items correctly, certain two-stage designs could perform as effectively as the "best" up-and-down procedures across the targeted ability range. Second, when low-ability examinees had approximately a 20% chance of answering difficult items correctly due to guessing, it was challenging to identify a two-stage procedure that matched the effectiveness of the "best" up-and-down procedures over the ability range of interest.

Key Limitations: Identifying truly "optimum" designs was not feasible; instead, the study relied on an iterative process of investigation and modification. Determining optimal cutting points on the routing test was difficult using trial-and-error methods. The scoring method employed did not possess strictly optimal properties, particularly for small total test length as the information function may not possess all its desirable asymptotic characteristics for 15-item tests. Moreover, the current study did not consider shorter tests in which guessing could occur.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test* (ETS Research Report No. RB-74-30). Princeton, NJ: Educational Testing Service.

Description: This study presented practical methods for redesigning a homogeneous test as a multilevel test and illustrated their applicability by examining potential modifications to the SAT. The study addressed questions regarding how the relative efficiency of an existing test might change when items are added, removed, or replaced. Using an existing SAT form, the study outlined key concepts of multilevel testing, including the routing test, equating to place different levels of multilevel tests on a common scale, the arrangement of levels with items shared between two adjacent levels, the relative efficiency of each level compared to the original SAT form, the dependence of the multilevel test on its levels, and procedures for assigning examinees to levels.

Key Findings: The results indicated that a multilevel test with three or four levels could achieve reasonably spaced difficulty levels. The study suggested that, as long as modifications remain within the scope of the existing homogeneous test, it is not necessary to build and test each possible design to predict the relative efficiency of a redesigned test.

Key Limitations: The findings were based on theoretical models and simulations, which may not fully capture the complexities of real-world testing scenarios. The study focuses on redesigning existing tests, as data for entirely new tests is often unavailable, and findings (e.g., replacing reading items) may not generalize to other scenarios. Furthermore, the author mentioned that the proposed methods would require careful consideration of various factors, such as item difficulty, test length, and examinee ability levels, which could be challenging to balance effectively. Furthermore, the study did not account for potential practical issues like test administration constraints or the impact of guessing on test outcomes.

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions* (Unpublished doctoral dissertation). The University of Maryland College Park.

Description: This study examined the impact of local item dependence (LID) of testlet items on making pass/fail decisions in three-stage MST. For simulation, the study manipulated conditions for test length (24 and 36 items), the proportion of testlet items (0, .33, .67, and 1), the position of testlet items, LID magnitude (.25, 1, and 1.5), and the measurement models. For the proportion of .33 and .67, testlet items were assigned to one or two stages, respectively. For measurement models, the study considered two models: (1) the 3PL unidimensional IRT (UIRT) model, in which ignores LID, and (2) the 3PL TRT model, which accounts for LID. For each measurement model, the study generated an item pool and assembled a 1-2-2 panel with the equal numbers of items per stage. Items for the medium and hard modules were selected to maximize information at $\theta = 0$ and 1, respectively. For each study condition, 100 examinees were simulated at each θ point from -3 to 3 in increments of .25. Responses were generated using the 3PL TRT model, and abilities were estimated under either measurement model. Cutscores for routing and pass/fail decision were set at .5 and 1, respectively. Results were evaluated in terms of bias and RMSE of ability estimates and decision accuracy (DA).
Key Findings: When the 3PL UIRT model was used, both ability estimation accuracy and DA declined as the LID level increased. Accuracy of ability estimates and DA also decreased with higher proportions of testlet items and shorter test lengths. Using the TRT model yielded only modest improvements in overall ability estimation accuracy and DA, suggesting that the UIRT model inflated precision in ability estimates and DA.The position of testlet items had minimal effect on outcomes.
Key Limitations: For the item pools, item parameters were estimated under the assumption that all examinees responded to all items. Also, when selecting items for the MST panels, the study considered only psychometric, ignoring other aspects such as content control, the proportion of testlet items per module, and the ability points at which maximum information occurred.

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Description: This paper described computer-adaptive sequential testing (CAST) within a multi-stage and testlet-based framework and discussed four aspects of CAST: panel design, automated test assembly (ATA), security control, and large-scale data management. In designing CAST panels, it is important to consider both test score precision and adherence to test specifications. To address concerns related to score precision, the article presented four strategies for generating appropriate test information functions (TIFs): the average maximum information technique, the middle-out strategy, the common-first-module strategy, and the separate-and-average-the-first strategy. With respect to ATA to build CAST panels, the article discussed the normalized weighted absolute deviations heuristic (NWADH) and its application to the construction of CAST panels. The article further emphasized that CAST could achieve pre-specified item exposure rates through the preconstruction of panels. Furthermore, the article argues that CAST could provide a systematic framework for managing large-scale data structures.
Key Findings: The article offered a descriptive overview of CAST, addressing its four key aspects. Readers are referred to the original article for further details.
Key Limitations: The author stated that CAST is not a universal solution for every testing program. Multi-stage techniques like CAST may remain largely conceptual with limited practical applicability unless ATA software is capable of fully managing the complexities involved in large-scale panel construction.

Luecht, R. M. (2003). *Exposure control using adaptive multi-stage item bundles.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Description: This study introduced a test development paradigm for testlet-based multistage adaptive testing (MST/MAT), referred to as the bundled MST (BMAT) framework. The framework was designed to address psychometric concerns such as reliability and item exposure, as well as test development requirements including content balancing. Within BMAT, a bundle consists of six components: (1) a series of bins holding testlet, (2) a testlet bank, (3) an item bank, (4) specification sets for test assembly, (5) design template, and (6) a routing table. Each bundle has multiple bins, and each bin contains testlets that are parallel in terms of statistical properties and test specifications so that they can be used interchangeably. Since each bin is associated with its own set of ATA specifications, testlets can be preassembled accordingly and assigned to the testlet bank for the corresponding bin and bundle. The study also described the use of a routing table for selecting an appropriate bin at subsequent stages and the use of a prescribed design template specifying the number of stages, the number of bins per stage, and the relationships among bins within and across stages. Additionally, the study discussed exposure control mechanisms considering expected route proportions, bin-wise statistical targets, and the number of testlets per bin, and provided ATA options for BMAT implementation.

Key Findings: The author positioned this paper as a prescriptive contribution for research and development. Within BMAT framework, real-time testing can be possible through three mechanisms: scoring algorithms, routing algorithms, and the random selection of a testlet within each bin. However, despite advances in ATA technology, several practical challenges remained at the time the study was conducted.

Key Limitations: The study noted practical issues that could only be addressed in future research with further advances in ATA technology.

Luecht, R. M., Brumfield, T., Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Description: This study presented a testlet-based MST using the Uniform CPA Examination as an illustrative example. For a 1-3-3 testlet-based MST design, the study examined factors that could affect measurement quality and score accuracy, including testlet size, the number of items per stage, and the size and characteristics of the item bank. To achieve a total test length of 60 items, the study proposed ten different specifications with testlet sizes ranging from 15 to 30 items. The study also applied two routing methods (AMI and DPI). Furthermore, the study outlined ATA procedures for constructing MST panels.

Key Findings: The total number of items required per panel could be reduced by decreasing the number of items in later stages and by controlling item overlap across panels and across modules within the same stage. This approach enables the construction of a larger number of panels while lowering item exposure rates.

Key Limitations: The study addressed issues to be solved, including determining the optimal number of stages and the appropriate range of testlet difficulty within each stage; establishing feasible statistical information targets that account for both psychometric quality and item exposure; investigating differences between scores and decisions derived from NC score routing versus IRT-based score routing; and evaluating the limitations of various ATA approaches for constructing multistage tests.

Luecht, R. M., Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Description: This study presented the conditional information targeting (CIT) strategy, designed to achieve multiple goals in generating feasible testlet TIF targets for MST designs: 1) controling the proportion of the population routed along various pathways within a panel; and 2) maximizing the informativeness of the targets while accounting for item quality, content requirements, and other test specifications. The CIT strategy utilized three ability points–left, right, and center posts–to compute provisional target TIFs. For simulation, this study used a 1-2 MST design. For routing, three center posts were considered: -.524, -.255, 0. For each routing condition, the study employed two conditions the left and right posts: 1) allowing both left and right posts to vary, and 2) fixing the right post at .60. Results were evaluated based on the intersect points of the final TIF targets and the amount of information associated with these points.
Key Findings: Based on the study results, the targets were better defined when one of the outer posts was held constant.
Key Limitations: First, the results may depend on the number of TIFs averaged to produce the provisional targets. Second, future research should explore the extension of the CIT strategy to multiple stages with more than two testlets per stage, with each testlet targeting a distinct difficulty level. Third, the CIT strategy applies only to the primary routes; it requires future investigation into its impact on auxiliary routing across various MST panel configurations.

Luecht, R. M., Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249.

Description: This paper presented a CAST approach for test construction and administration. The CAST approach could preserve the efficiency of traditional CAT while allowing addaptation for computer mastery testing. Using medical licensure tests, three examples were provided to demonstrate various CAST panel configurations and ATA strategies. The study considered three panel designs (1-3, 1-3-3, and 1-3-5) and investigated both "bottom-up" and "top-down" approaches to ATA strategies. For the three simple applications of CAST, the study demonstrated how different panel configurations and ATA strategies could be employed to achieve both statistical and content-related objectives.

Key Findings: CAST appeared to be a flexible and creative method, allowing test specialists to review panels and modules as part of a quality assurance process in the test development. According to the authors, CAST embodies a different philosophy for producing high quality computerized tests, rather than serving as direct alternative to CAT.

Key Limitations: The paper did not seek to address all aspects of CAST and noted that future research should address unresolved issues, such as defining an optimal panel once mathematical programming technologies become available for ATA.

Luo, X., Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243-263.

Description: This study proposed a new top-down design approach where design parameters for all allowed routes were specified at the test level, and their global optimality was determined using an advanced ATA algorithm. The new approach consisted of five sub-processes: 1) route mapping, 2) setting objectives, 3) setting constraints, 4) routing error control, and 5) test assembly. A simulation study was conducted by varying test length (24, 48, and 60 items), MST configurations (1-3, 1-2-2, and 1-3-3), and design approaches (two top-down MSTs [TD-MSTs] and three bottom-up MSTs [BU-MSTs]). The two TD-MSTs included designs with and without controlling routing error; and the three BU-MSTs included designs with equal constraints across stages or with additional constraints applied to either the initial or final stage. The item pool consisted of 500 items randomly retrieved from a large-scale licensure examination. For each study condition, 10,000 examinees were randomly simulated from $N(0, 1)$. The study used the maximum information rule to route examinees across stages and the MLE method to estimate both interim and final abilities. The five MSTs designs were evaluated in terms of route information functions (RIFs), RMSE in ability estimates, and route usage rate.
Key Findings: The TD-MSTs outperformed the BU-MSTs in four ways: (1) inproved separation and parallelism of RIFs, (2) enhanced control over the test-level characteristics, including RIFs, routing precision, and route usage rates, (3) greater measurement precision, and (4) the opportunity to review and refine design decisions. However, since the relationship between the TD and BU approaches were complementary, the study recommended the combined application of both methods in practice.
Key Limitations: This study used a single open-source software as the solver; using alternative solvers could deliver different results. Also, if the TD approach were applied in other contexts, such as licensure testing, the design process should emphasize priorities other than the measurement precision.

Luo, X., Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing*, *19*(3), 227-247.

Description: The study introduced a new adaptive testing design, dynamic MST (dy-MST) which incorporates both stage-level and item-level adaptation. In dy-MST, stage-level adaptation was added by constructing modules targeting different difficulty levels during the test development phase. During test administration, within each module, items were administered to examinees in decreasing order of information. When an examinee's projected ability range pointed the same module, the examinee was routed to the subsequent module even if not all items in the current module had been administered. For simulation, the study compared three testing designs: dy-MST, MST, and CAT. For MST and dy-MST, the study varied conditions for two factors: MST designs (1-2, 1-3, 1-4, 1-2-2, 1-2-3, and 1-3-3) and item partition strategies. For item partition strategies, three strategies were considered: equal-priority (EQP), first-stage-priority (FSP), and last-stage-priority (LSP) partitions. An item pool of 480 items was generated using 3PL item parameters drawn from known distributions. For MST, one panel was constructed for each study condition; and, cutscores were determined so that even proportion of the population could be routed across stages. For CAT, the most informative item from the shadow test was administered to each examinee. For each crossed study condition, 10,000 simulees were generated from $N(0, 1)$ and administered all three testing designs. For all designs, the maximum test length was set to 60 items. Performance was evaluated in terms of measurement precision, balance in content and response time, and changes in SEM throughout the test.

Key Findings: For dy-MST, test lengths were reduced to 63-74% of the maximum test length, while classification accuracy rates remained comparable to that of the full-length test and CAT. Based on the results, dy-MST seemed promising for delivering adaptive tests, even though dy-MST relied on preselected ("handpicked") items prior to administration. Furthermore, test lengths in dy-MST tended to decrease as the number of stages and modules per stage increased; the effect of the number of modules per stage was larger than that of the number of stages. Among the three item partition strategies, tests using the LSP partition were shorter than those using the FSP partition.

Key Limitations: The study did not examine item exposure control to the extent that may be required in operational testing; item exposure rate could depend on factors such as the testing program, item pool size, testing volume, security management, and examinee population characteristics. Moreover, the study considered only two content constraints–content distribution and speededness–which might be insufficient for real-world testing applications.

Ma, Y. (2020). *Investigating hybrid test designs in passage-based adaptive tests* (Unpublished doctoral dissertation). The University of Iowa.

Description: The study proposed a hybrid design for passage-based adaptive testing (HMCAT design). For a given MST configuration, the HMCAT design first assigned passages to modules while accounting for both statistical and non-statistical constraints. Within each passage, items were selected adaptively using the MFI method. Item-level CAT could also be implemented in a routing stage, a final stage, or across all stages. Accordingly, the study considered four HMCAT designs: (1) item-level CAT for all stages (CC), (2) item-level CAT in the first stage (CP), (3) item-level CAT in the last stage (PC), and (3) no item-level CAT (PP). For designs with more than two stages, item-level CAT was applied either in the first or the final stage. For item-level CAT, items with maximum weighted information at the examinee's estimated ability were selected. For simulation, the study considered two panel configurations (1-3, 1-3-3, and 1-3-3-3) and assigned four passages of five5 items each. The number of passages per stage varied according to the number of stages (1-3, 2-2, 1-1-2, or 1-1-1-1). Routing cutscores were determined using the DPI method, and abilities were estimated using the EAP method. For each study condition, 5,000 simulees were generated from $N(0, 1)$ with 50 replications. Performance was evaluated in terms of the accuracy of ability estimates.
Key Findings: The performance of the HMCAT designs varied MST configurations. Overall, the PC design outperformed the other designs under the 1-3-3 and 1-3-3-3 configurations. The CC and PP designs exhibited similar performance, while the CP design showed comparable results to the CC and PP designs for the 1-3 configuration. Additionally, under the 1-3-3 configuration, misrouting tended to deteriorate the performance of the CP design execlusively.
Key Limitations: The study employed the DPI method for routing and examined only a limited set of conditions for each study factor. Consequently, the findings should be interpreted within the specific scope of this study. Future research should explore alternative routing methods and additional factors.

Macken-Ruiz, C. (2008). *A comparison of multistage and computerized adaptive tests based on the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: The study compared the performance of MST and CAT under the GPCM. For simulation, the study considered three 1-3-3 designs that varied in the number of items per stage; the number of items per stage increased (3-5-12), decreased (10-7-3), and was equal across the three stages (6-6-6). Items were carefully selected from an item pool of a large national testing program to achieve content balance and to control item exposure rates. For routing, the study used rules based on maximum information, fixed θ, and number-right scores. Under the maximum information routing rule, examinees were routed to the module providing the most information at their current ability estimates. For the fixed θ routing rule, cutscores of -1 and 1 were used to route examinees to easy, moderate, and difficult modules. For the number-right routing rule, cutscores were determined so that the lower and upper thirds of score distributions were routed to the easy and difficult modules, respectively. Final ability estimates were obtained using the MLE method. For each study condition, the study conducted ten replications. For CAT, the test length was fixed to 20 items and abilities were also estimated using the MLE method. Performance of MST and CAT was compared with respect to ability recovery (bias, RMSE, and MAD) and item exposure rates.
Key Findings: Based on item exposure rates, the item pool was utilized more efficiently in CAT than in MST. Regarding ability recovery, the MST design with increasing items per stage exhibited the best performance, while the equal-items design performed the worst; however, the differences were minimal and overal performance was comparable across the three MST designs. The MST designs tended to provide more precise estimates in the middle of the ability distribution than CAT, whereas CAT produced slightly more precise estimates at the extremes. However, the differences were not substantial, indicating that MST could serve as a viable alternative to CAT. Among the three routing rules, the maximum information rule performed the best, followed by the fixed θ and the number-right routing rules.
Key Limitations: The study did not use the same number of items across the three MST designs. Moreover, since only a single panel was constructed for each MST design, the total number of items differed between CAT (208 items) and MST (three sets of 20 items). Furthermore, panel construction followed a top-down approach, in which items were selected to match TIFs rather than module-level information functions.

Martin, A. J., Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27.

Description: The study compared the implications of multistage adaptive testing (MST/MAT) and fixed-order computer testing in predicting various test-relevant outcomes. Using NAPLAN tests, the study calibrated items using the Rasch model and constructed four 1-2-3 MST forms, one for each of four year-levels (3, 5, 7, and 9). Correspondingly, four fixed forms were also assembled, each aligned with one of the test pathways available in the adaptive testing condition. To administer MST and fixed test forms, the study recruited 231 Australian school, with students within each school randomly assigned to one of the two testing conditions. At the end of testing, students completed brief surveys assessing test-relevant motivation and engagement, as well as subjective test experience. To predict outcome variables–including NAPLAN score, test-relevant motivation and engagement, and subjective test experience–the study fitted a two-level model with student-level covariates (test condition, gender, year-group) at Level 1, and school-level covariates (socio-educational advantage, location, structure, and size) at Level 2. The analysis focused on the main effects and interaction effects involving test condition.

Key Findings: The study found that measurement precision was greater for MST than for fixed forms, with the effect more pronounced for female and older students. Furthermore, the study results indicated that MST did not reduce test-relevant motivation, engagement, or subjective experience of students. The study also suggested that MST resulted in larger positive effects for older students, particularly at developmental stages with relatively lower motivation and engagement. The results indicated that taking MST could reduce achievement error rates and enhance test-relevant motivation and engagement. Additionally, older and female students were likely to achieve higher scores, exhibit greater motivation and engagement, and report more positive subjective test experience under the MST condition.

Key Limitations: Since the outcome variables for test-relevant motivation, engagement, and subjective test experience were self-reported, the accuracy of the data may have been compromised by low motivation or engagement in responding to survey items. Furthermore, since the study relied on data from a test administered by a central authority, only a limited subset of the full dataset was available for analysis. Future research should consider additional factors, including students with special needs, domains other than numeracy, and the P&P testing condition.

Martin-Fernandez, M., Ponsoda, V., Díaz, J., Shih, P.-C., Revuelta, J. (2016). A multistage adaptive test of fluid intelligence. *Psicothema*, *28*(3), 346-352.

Description: This research compared different MST structures within the context of the fluid intelligence test. The first study developed an initial item pool for the operational fluid intelligence MST and assessed its psychometric properties. Based on the operational item pool from the first study, the second study employed ATA to construct 1-2-2 and 1-3-3 MST configurations. To maximize information for high-ability examinees, the study allowed selecting identical items across modules in different stages and added five more items to the very difficult modules at the third stage. The appropriateness of MST configurations was assessed based on the number of actual paths and the amount of test information provided by these paths.

Key Findings: Across all paths, the average test information over 121 ability points from -3 to 3 was higher for the 1-2-2 configuration than for the 1-3-3 configuration. Within the ability range considered in this study, the 1-2-2 configuration provided the greater information when five, six, and four items were assigned to the first, second, and third stages, respectively (with nine items in the high-difficulty modules). However, the 1-3-3 configuration consumed the most informative items in the beginning stages and failed to achieve the same level of test information as the 1-2-2 configuration. Furthermore, for the 1-3-3 configuration, the study identified some unused paths and found that some items were administered more than twice within the same path.

Key Limitations: Since the study added five more items to the very difficult modules, module lengths were unequal at the third stage. Future research could consider enforcing equal module lengths within the same stage. Additionally, future studies might the dimensionality of paths in MST structures, gather predictive validity evidence, and explore models other than the graded response model.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185-187.

Description: This article discussed how MST can address practical limitations of CAT. In assembling tests, "on-the-fly" ATA for item-level adaptive CAT may require relaxing some specifications (e.g., content), potentially resulting in suboptimal test forms. However, MST can mitigate this issue by allowing subject-matter experts to review tests prior to administrations. During testing, MST permits examinees to skip, review, and modify their responses, features not allowed in CAT. From a statistical perspective, the item response matrix in CAT is sparse, limiting the applicability of traditional analyses. However, the block-sparse response matrix in MST supports a wider range of statistical analyses. Additionally, MST enhances test security by enabling control over item exposure rates.

Furthermore, the article provided a brief overview of four studies published in the same special issue. According to Mead (2006), the first article by Luecht, Brumfield, and Breithaupt (2006) described common practical aspects of MST and discussed an ATA method. The second article, Jodoin, Zenisky, and Hambleton (2006), compared the performance of several MST configurations and fixed-length tests. The study showed that measurement accuracy of MSTs were comparable to that of the fixed-length tests. The third article, by Hambleton and Xing (2006), examined additional MST designs and compared the performance of different testing administrations (MST vs. linear form vs. item-level adaptive). The final study, by Chuah, Drasgow, and Luecht (2006), investigated the effects of sample size on the accuracy of ability estimates and licensure classification.

Key Findings: The article provided a brief overview of four studies published in the same special issue. Readers are referred to the original article for further details.

Key Limitations: -

Park, R. (2013). *The impact of statistical constraints on classification accuracy for multistage tests.* Durham, NC: American Institute of CPAs.

Description: This study investigates the impact of statistical constraints on the classification accuracy of a 1-2-2 MST. The study manipulated levels of TIF (40-100% in increments of 10% compared to the maximum value), overall test difficulty (easy, medium, and hard levels), and cutoff scores (θ = -0.5, -0.3, -0.1, 0.1, 0.3, and 0.5). To construct MST panels, the study modified an item pool obtained from an operational licensure exam; the pool consisted of 3,046 dichotomously-scored items whose parameters were calibrated using the 3PL IRT model. For each MST panel, each module included 25 items selected to achieve target TIFs at specified ability points. Panels were constructed using R (R development Core Team, 2013) and the lpSolveAPI package (Konis, 2011). For each crossed study condition, the study performed one hundred replications, each with 1,000 examinees whose abilities were randomly generated from $N(0, 1)$ or a uniform distribution in (-4, 4). At the end of each stage, abilities were estimated using the EAP method. Performance was evaluated in terms of correct classification rate (CCR), false positive error rate, and false negative error rate.

Key Findings: The study found that classification accuracy increased as the distance between test difficulty and cutoff scores became narrower. Additionally, CCR decreased as test information declined; a drop in CCR seemed significant (i.e., more than a 2% drop) when TIF decreased below 60% of the maximum value considered in the study. However, the rate of increase in CCR decreased as the test information increased, suggesting that there was "diminishing return" between test information and test accuracy in classification.

Key Limitations: Evaluating target TIFs prior to test construction poses a significant challenge for test developers. Moreover, when multiple panels are required for operational testing, it is unlikely that all panels will achieve the maximum test information. The author also noted that conventional methods for determining TIFs are not formally validated to guarantee optimal TIFs.

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* (Unpublished doctoral dissertation). The University of Texas at Austin.

Description: The study evaluated the measurement accuracy of mixed-format MST designs. For simulation, the study varied conditions for total score points (40 and 60), panel designs (1-2-2 and 1-3-3), the proportion of polytomous test units in routing modules (purely dichotomous and mixed-format), and the proportion of polytomous test units in the test (10%, 30%, 50%, and 70%). To construct MST panels, the study used a mixed-format test unit pool from the NAEP science and applied ATA techniques to build panels to achieve target TIFs for each pathway. For each study condition, the study conducted 100 replications, each with 1,000 abilities generated from $N(0, 1)$. Abilities were estimated using the EAP method; and routing was based on examinees' estimated ability relative to pathway difficulties (θ = -1.2, 0, and 1.2 for the 1-3-3 design and -.5 and .5 for the 1-2-2 design). Throughout the whole simulation procedure, the study used the GPCM. Results were evaluated in terms of ability estimate accuracy and test information functions.

Key Findings: For the small and medium proportions of polytomous test units (i.e., 10%, 30%, and 50%), results were similar. Ability estimation accuracy improved as information increased. Although both MST designs exhibited comparable standard errors near the center of the ability distribution, the 1-3-3 design tended to yield smaller standard errors across a wider range of abilities than the 1-2-2 design. Moreover, results were similar across the two routing module designs. When the proportion of the polytomous test units was as large as 70%, standard errors for low-ability examinees were higher than those observed for smaller proportions. Additionally, at 70%, the effect of the routing module design more pronounced. When the routing module contained only dichotomous items, standard errors were larger–particularly for examinees with low abilities–compared to those of the mixed routing module design. Measurement accuracy improved as the total test points increased.

Key Limitations: The test unit pool did not have enough easy test units, which made it challenging to construct MST panels with a high proportion of polytomous test units. The current study focused exclusively on measurement accuracy only and did not account for factors such as test-taking time and the development costs associated with polytomous test units. Therefore, future research should consider additional factors, including test-taking time, content balancing, and scoring of constructed responses, which may affect results.

Park, R., Kim, J., Chung, H., Dodd, B. G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. *Applied Psychological Measurement*, *38*(4), 268-280.

Description: This study examined a new method of reassembling MST panels to enhance pool utilization in constructing mixed-format MST based on the GPCM. The reassembly procedure allowed previously utilized test units to be replaced with unused items. Building on the initial MST construction, MST panels were reassembled using a linear programming (LP) model. A simulation study was conducted by varying levels of reassembly (first, second, and third), test unit replacement ratios (TRRs: 0.22, 0.44, and 0.66), and passing rates (30%, 50%, and 70%). The program JPLEX was used to assemble panels and modules so that the target TIFs could be achieved at specified ability points; and, the IRTGEN SAS macro was used to generate 1,000 normally distributed simulees across 100 replications. Abilities were estimated using the MLE method and routing was completed based on the AMI method. Results were evaluated in terms of statistical properties of MST modules (e.g., RMSEs in TIF), classification accuracy, and overall pool utilization.

Key Findings: Results showed that MST reassembly improved overall pool utilization while maintaining the desired MST construction. Moreover, all MST testing conditions demonstrated comparable classification accuracy. Additionally, the RMSE tended to increase as TRRs became higher.

Key Limitations: Future research should examine both the accuracy of ability estimates and the precision of item parameter estimates. Additionally, future studies should consider other factors, such as panel designs and LP modeling methods, to evaluate the feasibility of the proposed method in operational testing contexts.

Park, R., Kim, J., Chung, H., Dodd, B. G. (2017). The development of MST test information for the prediction of test performances. *Educational and Psychological Measurement*, *77*(4), 570-586.

Description: This study proposed a new method for predicting the performance of MST without conducting simulations. First, MST test information was derived using the active part of information from each primary pathway. For each θ point, test information was finalized by averaging the values over five points centered around θ. Measurement precision was then assessed by computing the analytic standard error at each θ_i ($SE(\theta_i)$), defined as the inverse of its test information. For classification accuracy, false positive error rates (FPER) and false negative error rates (FNER) were obtained under the assumption that $\hat{\theta}_i$ follows a normal distribution with mean θ_i and standard deviation $SE(\theta_i)$. To evaluate the new method, the study conducted simulations varying test length (40 and 60 total points), MST panel structure (1-2-2 and 1-3-3), routing module design (dichotomous items only and mixture of dichotomous and polytomous items), and the proportion of polytomous items (10%, 30%, 50%, and 70%). To determine pass/fail for examinees, the study used three cutscores (-.524, 0, and .524) corresponding to 30%, 50%, and 70% passing rates, respectively. For each condition, the study generated 1,000 examinees from $N(0, 1)$ with 100 replications. Abilities were estimated using the EAP method and routing was based on the AMI method. Simulation results were compared with the analytic results obtained from the new method in terms of conditional standard errors and classification accuracy.

Key Findings: Study results showed that the performances of the new method were very close to the simulation results, suggesting that MST performances–both measurement precision and classification accuracy–could be directly predicted from test assembly results.

Key Limitations: The authors did not explicitly acknowledge the limitations of their study.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Description: This study compared the performances of P&P, CAT, and MST with respect to ability estimation accuracy (bias and RMSE), the relative efficiency of MST designs, and item exposure rates. Test length was fixed at 36 items across all test designs. For simulation, the study selected items from an existing item pool of the LSAT Logical Reasoning section. For the P&P design, a single form was constructed based on a target TIF representing a typical CAT. In the CAT design, subsequent items were selected using the maximum information method, with a maximum conditional exposure rate of 0.25. For MST, the study constructed 12 panels by varying the number of stages (2 and 3), the number of subtests/modules per stage (3 and 5), and the number of items per module. For the number of items per subtest, the study considered three conditions: increasing, equal, and decreasing numbers cross stages. For each study condition, ability points ranged from -2.25 to 2.24 in increments of 0.5, with 500 replications at each point. Throughout the simulations, the study incorporated the 3PL IRT model and abilities were estimated using the MLE method.

Key Findings: Based on the study results, the CAT design showed higher accuracy in ability estimates and lower item exposure rates compared to the MST designs. However, the study noted that advantages of MST could make MST an attractive alternative to CAT. Among the 12 MST designs, ability estimation accuracy increased as the number of stages increased. Both ability estimation accuracy and the efficiency of the MST designs also increased as the number of subtests increased, particularly at ability points between -0.75 and 2.25. However, the number of items per stage had only a minimal effect.

Key Limitations: Given that the study considered only a limited range of factors and conditions for MST designs, the results should not generalized to broader contexts.

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, *50*(4), 447-468.

Description: The study proposed a new method for applying adaptive testing in longitudinal large-scale studies called longitudinal multistage testing (lMST), and compared its performance to that of conventional testing (CT; i.e., P&P mode). To evaluate the performance of lMST, the study applied the design to data from National Educational Panel Study (NEPS) using the Rasch model. In the CT design, all examinees received the same test form of 35 items. For adaptivity in lMST, test forms were constructed at three difficulty levels (easy, medium, and hard). The study varied both the characteristics of link items across test forms (broad vs. narrow difficulty range) and the number of link items (strong vs. weak). In lMST, approximately 56% of examinees were routed to the medium form, while about 22% were assigned to either the easy or hard form. The three lMST test forms were used to investigate how form difficulty and the number of link items affected ability estimation accuracy, change scores, and misallocation rates. For each study condition, the study conducted 100 replications.

Key Findings: lMST outperformed CT in test targeting, which could possibly enhance student engagement and reduce missing data. Across the full ability range, ability estimates were more precise for all lMST designs than those for CT. Both CT and lMST showed relatively high standard errors, suggesting that neither may be suitable for providing individual-level feedback. Regarding the precision of change score estimates, lMST performed better when correlations between testing waves were high, whereas CT had better results when correlations were low.

Key Limitations: This study was limited to a single sample size, one set of cut points, one fixed number of test forms, a specific set of item difficulty distribution for the three test forms, and a single IRT model. Future research could expand study conditions to explore a broader range of scenarios. Furthermore, since the effects of missing responses and test length could vary across lMST designs, future studies should also consider these factors.

Reese, L. M., Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing. law school admission council computerized testing report.* (LSAC Research Report No. LSAC-R-96-04). Princeton, NJ: Law School Admission Council.

Description: This study compared the performance of 1-3 MST designs with a standard fixed-length CAT and a P&P form. The two-staged MST administered a total of 25 items, with two testlets of 10 items at Stage 1 and three testlets of 15 items at Stage 2. Examinees were routed to subsequent modules based on NC scores to minimize mean squared error of the NC scores at Stage 2. At the end of testing, final abilities were estimated using Bayesian modal scoring. For the fixed-length CAT, all examinees received 25 items, with each item randomly selected from the top items providing the highest information at the current ability estimate. For the P&P design, the study considered the test lengths of 25 and 50 items. Across all three designs, 7,000 simulees were generated from from $N(0, 1)$. Item responses were modeled using the 3PL IRT model. Results were evaluated in terms of ability estimation accuracy using bias and RMSE.

Key Findings: Based on the results, the study concluded that, when testlets and modules are carefully assembled, a two-stage testlet-based MST can achieve higher accuracy in ability estimates in the middle of the ability range compared to the other two designs. In fact, ability estimates for the two-stage testlet design were more precise than those for the doubled-length P&P design.

Key Limitations: The MST design did not allow misclassified examinees to change levels in the second-stage. The study suggested that allowing such changes could improve the accuracy of ability estimates, even at the extreme ends of the ability range.

Rome, L. (2017). *Evaluating item selection methods for adaptive tests with complex content constraints* (Unpublished doctoral dissertation). The University of Wisconsin-Milwaukee.

Description: The study compared the performances of CAT, MST, and OMST. A Monte Carlo simulation study was conducted to generate item pools under varying pool size (360 vs. 720 items), complexity of content constraints (basic, simple, medium, and complex), and representation of each content category (realistic vs. equal proportions). For each pool, MST designs varied by panel structure (1-3-3 vs. 1-4-4) and the number of items per stage (equal, decreasing, or increasing), while OMST and CAT varied by item selection method (maximum priority index [MPI] vs. weighted penalty model [WPM]). MST modules were constructed using a bottom-up and backward assembly approach. For the pool of 360(720) items, 4(8) and 3(6) parallel panels were constructed for the 1-3-3 and 1-4-4 designs, respectively. Each condition included 1,000 examinees sampled from $N(0, 1)$ and was replicated 100 times. Across all designs, test length was fixed at 36 items, and abilities were estimated using the EAP method. Results were evaluated in terms of content coverage (average number of constraint violations per test, deviation from test specifications), measurement precision (RMSE and bias), and item exposure and test overlap rates.

Key Findings: Overall, CAT and OMST outperformed MST in terms of content alignment and measurement precision. As content constraints became more complex, content coverage declined for OMST and CAT, although test security improved. In contrast, the effects of constraint complexity were minimal for MST. Regarding test security, CAT outperformed MST and OMST. For all designs, larger item pools increased test security, with the effect most pronounced for MST. Accuracy of ability estimates also improved with larger pools, although the effect was minimal for MST. For the 1-3-3 MST, measurement precision increased when more items were assigned to later stages. Furthermore, item exposure rates for OMST and MST became worse as the first-stage module length increased. For CAT and OMST, the=WPM yielded better content alignment and measurement precision than the MPI.

Key Limitations: The item pools did not reflect realistic relationship between item parameters and content categories. Also, the content weights for the item selection methods remained constant across different item pools and test blueprints. In practice, these weights could be adjusted in the MPI to achieve better content coverage.

Rotou, O., Patsula, L., Steffen, M., Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* (ETS Research Report No. RR-07-04). Princeton, NJ: Educational Testing Service.

Description: This study examined the performance of MST in comparison with CAT and P&P. Specifically, a 1-3 MST design with 33 items was compared to a 32-item CAT, and a 1-3 MST design with 54 items was compared to a 55-item P&P test. Comparisons were conducted using the one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL) based on item sets. The item pool consisted of 440 items from eight forms of the MCAT verbal reasoning test, with item parameters estimated under each IRT models. For MST,two parallel forms per module were assembled to satisfy content constraints and shape of information functions. For each IRT model, the study generated abilities for 500 examinees at NC scores from 16 to 54 in increments of two. In the MST designs, abilities were estimated using the MLE method and examinees were routed to the modules whose difficulty best matched their estimated abilities. For CAT, abilities were also estimated using MLE, with items selected to maximize information at the examinee's current ability estimate while controlling item exposure rates. Results were evaluated in terms of reliability coefficients and the conditional standard error of measurement (CSEM) at each generated number-correct score.

Key Findings: Study results showed that, for the 2PL and 3PL models, MST achieved higher reliability coefficients and lower CSEMs than an equivalent-length P&P test. Furthermore, for the 1PL and 2PL models, MST outperformed an equivalent-length CAT test. For the 3PL model, both MST and CAT exhibited comparable reliability.

Key Limitations: Since the items were set-based, the local independence assumption may have been violated, potentially inflating reliability estimates and the test information function. Additionally, the item pools for each IRT model differed slightly due to the removal of poorly fitting items or to ensure greater homogeneity in set difficulty. The study also noted that, since items were set-based, the accuracy of ability estimation and efficiency in CAT could decrease if an inappropriate item is selected early in the test.

Sadeghi, K., Khonbi, Z. A. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language Testing in Asia*, *7*.

Description: This paper provided an overview of DIF within the context of multistage computer adaptive testing (MSCAT) utilizing the 3PL IRT model, and offered practical recommendations for its implementation. The paper also discussed fundamental concepts and rationales underlying MSCAT. The paper proposed a model for identifying DIF in MSCAT and conducted a simulation study to demonstrate its application. The simulation used four-item modules within a seven-stage panel design. The first stage consisted of a single module, while stages two through seven each included three modules, each with items targeting at different difficulty levels. The study employed the NC-score routing strategy. The DIF items were generated from a normal distribution with the specified mean and standard deviation.
Key Findings: The program developed for DIF analysis in CAT (CATSIB) accurately and consistently detected DIF in an MSCAT of English proficiency across different focal and reference groups using the 3PL IRT model.
Key Limitations: The authors did not explicitly mention the limitations of their study.

Sari, H. I., Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory and Practice*, *17*(5), 1759-1781.

Description: The study compared the performances of CAT and two MST designs by manipulating the number of content areas and test length. For simulation, the study varied administration mode (CAT vs. MST), test length (24 vs. 48 items), and the number of content areas (zero for no content control, 2, 4, 6, and 8). The study used an item bank from a real ACT math test that included 480 items across six content areas, with item parameters estimated using the 3PL IRT model. Four additional item banks were constructed to reflect the varying number of content areas. For the 1-3 and 1-3-3 MST designs, items were selected for modules to maximize test information at fixed θ points; multiple parallel panels were constructed to control item exposure rates. For each condition, 4,000 simulees were generated from $N(0, 1)$ with 100 replications. Abilities were estimated using the EAP method. And, the maximum information method was used for item selection in CAT and routing in MST. Accuracy of ability estimates were evaluated using overall statistics (mean bias, RMSE, and corr(θ, $\hat{\theta}$)) and conditional statistics (CSEM).

Key Findings: The number of content areas had minimal impact on the performance of both CAT and MST designs, whereas test length and administration mode had more pronounced effects. Across all study conditions, CAT outperformed the two MST designs, and the two MST designs exhibited comparable performance to each other.

Key Limitations: Test length was held constant for both CAT and MST designs. If measurement accuracy were used as a stopping rule in CAT, the resulting measurement precision would be similar for CAT and MST. The study also employed unrealistic simulation settings. When selecting the two content areas, the study chose one easy and one hard area to avoid confounding due to content difficulty, and kept the same number of items across content areas.

Sari, H. I., Raborn, A. (2018). What information works best?: A comparison of routing methods. *Applied Psychological Measurement*, *42*(6), 499-515.

Description: The study compared the performance of five routing methods within the ca-MST framework. The methods were: MFI, maximum likelihood weighted information (MLWI), maximum posterior weighted information (MPWI), Kullback-Leibler (KL), and posterior KL (KLP) methods. For simulation, the study varied test length (30 vs. 60 items) and panel design (1-3, 1-2-2, 1-2-3, and 1-3-3). For each module, items were selected from a simulated item bank to maximize the information function at a fixed θ point. Fixed points were set at 0 for routing modules. For stages with two modules, fixed points were set at -1 and 1 for easy and hard modules, respectively. For stages with three modules, easy, medium, and hard modules had fixed points of -.5, 0, and .5, respectively. Across all designs, the study constructed three non-overlapping parallel panels. For each condition, 3,000 examinees were simulated from $N(0, 1)$ and randomly assigned to one of the parallel panels; this procedure was repeated 100 times. Results were evaluated using both overall (mean bias, RMSE, corr(θ, $\hat{\theta}$), and module exposure rates) and conditional (CSEM, absolute bias, and RMSE) statistics.

Key Findings: Overall and conditional outcomes improved as test length increased. The effects of routing methods and panel designs on overall results were minimal. However, for the conditional outcomes, panel design had a noticeable impact, primarily because fewer stages resulted in more items per module. For a test length of 30 items, the study results recommended the 1-3 design with the exception when using the MFI method. For 60-item tests, the 1-3-3 design provided stable results across all routing methods. Among other 60-item designs, the MFI, KL, and MLWI methods performed best, with MLWI producing consistently low bias and RMSE.

Key Limitations: The study examined only a limited set of conditions for the study factors. Future research should explore additional conditions for routing methods, panel designs, and panel assembly techniques. It should also compare the performance of CAT and ca-MST.

Sari, H. I., Sari, H. Y., Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406.

Description: To familiarize researchers—particularly those in Turkey—with ca-MST, this article provides an overview of the framework. The article is organized as follows:

- Introduction to Test Administration Models: Discusses P&P, CAT, and MST designs, providing examples for each.
- Structure and Working Principle Of Ca-Multistage Testing: Explains how MST can be administered.
- Building A Computer Adaptive Multistage Test: Describes MST components related to design, routing, and scoring, and reviews related research.
- On The Fly Computer Adaptive Multistage Testing: Provides a brief overview of on-the-fly ca-MST.
- Books and Software for Computer Adaptive Multistage Testing
- Discussion and Future Research on ca-MST

Key Findings: -
Key Limitations: -

Schnipke, D. L., Reese, L. M. (1997). *A comparison of testlet-based test designs for computerized adaptive testing.* A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Description: This study incorporated testlets into MST designs and compared the performances of MST, CAT, and P&P test designs in terms of ability estimation accuracy. The study considered a 1-3 MST panel design, with two testlets assigned to routing modules. Within the second stage, examinees received either the same set of three testlets or a set of three different testlets, depending on their NC scores from previous testlets. As a variation to the two-stage MST with adaptation at testlet-level, the study also considered a 1-3-4-5 design, in which one testlet was assigned to each module of the second, third, and fourth stages. As a result, across all designs, each examinee completed a total of five testlets, with a maximum possible NC score of 25. For the P&P design, the study considered test lengths of 25 and 51 items selected from LSAT test sections. In CAT, subsequent items were selected to maximize information at the examinee's current ability estimate, with a fixed test length of 25 items. For each study condition, 1,000 simulees were generated at each θ point from -3 to 3 in increments of 0.25, yielding a total of 25,000 examinees. In MST, examinees were routed to modules or testlets that minimized the MSE of their NC scores. Across all designs, final abilities were estimated using Bayes modal scoring. Results were evaluated in terms of ability estimation accuracy, using bias and RMSE.

Key Findings: In terms of both bias and RMSE, CAT outperformed the other designs, particularly at the extremes of the ability distribution. Ability estimates for the MST designs were more precise than those for the 25-item P&P test across all ability points; the performances of the MST designs were comparable. When the P&P test length was doubled to 51 items, both RMSE and bias were similar to or better than those for the MST designs. These results suggest that MST designs are feasible options for computer-administered tests.

Key Limitations: The authors did not explicitly mention the limitations of their study.

Stark, S., Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education*, *19*(3), 257-260.

Description: The article first described the advantages of MST that 1) it could provide better test security; 2) it would allow greater control over test construction; 3) items could be skipped, and answers could be reviewed or changed within each module for test takers; and, 4) problems with sparse operational data matrices could be mitigated. Furthermore, the article introduced four articles in the special issue. Those articles provided some interesting insights about MST and the expected gains in reliability and decision accuracy.

Key Findings: Leucht et al. (2006) described the steps to construct 1-3-3 MST panels via ATA and present two alternatives (the AMI and DPI methods) for setting the upper and lower bounds on NC scores used for routing examinees through each panel. Jodoin et al. (2006) described how MST would lie between LFT and CAT in terms of administrative control and testing efficiency. They emphasized having reliable classification decisions than just reliable ability estimates. Hambleton and Xing (2006) compared the effects of optimal/nonoptimal test designs on decision accuracy by matching the target information function for panels to either the distribution mean of examinee abilities or to the passing score. Chuah, Drasgow, and Luecht (2006) discussed the cost associated with pretesting items for the launch of an MST program.

Key Limitations: Initial studies suggest that, for 40- to 60-item credentialing exams, the psychometric gains of MST over LFT are small, as test length could be the main factor improving accuracy in ability estimation and classification. Furthermore, the article points out the need for more research to understand the sources of misclassification decisions (e.g., issues with automated test assembly panels, problematic routes, or differences in examinee trait distributions). Consequently, if MST is not applied in specific contexts requiring much shorter tests (e.g., 15–20 items or fewer per dimension), its implementation costs may need to be justified primarily by concerns about test security and examinee satisfaction.

Svetina, D., Liaw, Y.-L., Rutkowski, L., Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192-213.

Description: The study examined the effects of factors related MST designs and routing examinees. For simulation, the study employed a 1-2-3 MST design with 36 items and varied the number of items per testlet, routing method, and routing probabilities. For the number of items per testlet, four configurations were considered: (1) equal (12-12-12), (2) short-to-long (6-10-12), (3) long-to-short (12-10-6), and (4) short-long-short (6-20-10). For routing methods, the study considered five approaches: (1) random selection, (2) NC cumulative score from all previous stages, (3) NC score from the previous stage, (4) IRT EAP estimate, and (5) IRT MFI function. Routing probabilities to optimal testlets were set at 1.00, .80, and .70, with 1.00 considered optimal routing and the other two as suboptimal. For each study condition, 4,000 simulees were generated from $N(0, 1)$ with 100 replications. Results were evaluated in terms of the recovery of ability and item parameter estimates, as well as item exposure rates.

Key Findings: With respect to the recovery of ability estimates, IRT scoring outperformed NC scoring. However, given the small differences in bias, NC scoring may be more practical due to its simpler algorithm. Assigning a smaller number of items to the first stage also tended to improve ability estimates. With respect to the recovery of item parameter estimates, both difficulty and discrimination parameters were estimated accurately, particularly for moderately difficult items. Additionally, item exposure rates were more evenly distributed under suboptimal routing compared to optimal routing.

Key Limitations: The scope of this study was limited. Future research should extend simulation studies by including other possible scenarios, such as multiple groups, violations of the measurement invariance assumption, or the presence of item drift, to enable more comprehensive conclusions.

Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* (Unpublished doctoral dissertation). The University of Illinois at Urbana-Champaign.

Description: This study proposed two new OMST approaches and one new hybrid design. The first OMST approach selected items over a wider ability range in the early stages and gradually narrowed the range as testing progressed. The second approach increased the chance of selecting items that provide greater information at an examinee's true ability by considering both provisional ability estimate and item information at relevant ability points. The hybrid design combined an OMST step in the initial stages with a transition to CAT in subsequent stages. For the first approach, three simulation studies were conducted: (1) to determine the optimal ability range at the first stage for each of the proposed OMST, (2) to compare the performances of the three proposed item selection designs with that of Zheng and Chang (2014), and (3) to evaluate the best design identified in the second study. For the second approach, two simulation studies were conducted to evaluate the performances of OMST-Uniform with Weighting Method Design (OMST-UWM) and the OMST-Uniform (OMST-U) design in terms of ability estimation accuracy as well as the classification accuracy and consistency for mastery testing. For the hybrid design, three studies were conducted to replicate the work of Wang, Lin, Chang, and Douglas (2014) and compared the new hybrid design with existing CAT designs.
Key Findings: Overall, the results indicated that the new designs could address limitations of CAT, such as underestimation and overestimation of ability, while ensuring higher test security compared to existing MST and CAT designs. Nevertheless, the optimal design should be selected based on the specific goals of the researchers.
Key Limitations: An unrealistic violation constraint was applied in all simulation studies. Additionally, all item parameters were simulated rather than drawn from an actual item bank.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117-130.

Description: The study demonstrated the use of a probabilistic response-time model to detect differential speededness in MST and to estimate differences in time intensity and speed across subtests and test takers. The authors proposed a six-step analysis procedure for diagnosing a test for differential speededness. To illustrate the procedure, the study used an empirical MST data set from the computerized CPA Exam (1-2-2 panel design). The data set consisted of responses for 1,104 examinees to a single subject test with 96 MC items. Throughout the analyses, the study employed the 3PL IRT model.
Key Findings: The study found that, although more difficult subtests contained items that were generally more time-intensive than those in easier subtests, the analysis of residual response times did not reveal any significant differential speededness. Additionally, the results indicated minor but consistent patterns of positive residual log times at the beginning of each subtest, likely refleecting a warm-up effect in which examinees spent more time on initial items than necessary.
Key Limitations: It was challenging to empirically verify the assumptions that, for a given item, a test taker's responses and response times are locally independent.

Wang, C., Chen, P., Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3-28.

Description: Since MST violates the missing-at-random (MAR) assumption, calibrating items using incomplete data can be challenging, especially when a test includes items measuring different but correlated subscales or subcontents. The study examined three calibration methods: (1) the MML estimation method, (2) the expectation maximization (EM) method, and (3) the fixed parameter calibration (FPC) method, applied to either the single group (SG) or the multiple groups (MG). For calibrating items, the study considered two scenarios: (1) all items calibrated on a single scale and (2) items from different subscales calibrated separately on distinct scales. For the second scenario, the study proposed a new approach based on Rubin's (1976) missing data theory, which satisfied the MAR assumption by augmenting subscale data. The study conducted two simulation studies, Design I and Design II, in which routing was based on true θ and estimated $\hat{\theta}$, respectively. The performance of the calibration methods was evaluated under a 1-3 MST design. Furthermore, the study used real data from NAEP and compared the effectiveness of the different approaches in estimating item parameters.

Key Findings: When all items were calibrated on a single scale, the performances of the MML and FPC methods were similar across all conditions. When examinees were routed based on their true θ–thereby violating the MAR assumption–the multiple-group (MG) approach produced more accurate item parameter estimates than the single-group (SG) approach, which resulted in severely biased estimates. On the contrary, when $\hat{\theta}$ was used for routing, the SG approach outperformed the MG approach. When items were calibrated separately by subscale, the MG approach outperformed the SG approach. Furthermore, the proposed calibration method improved the accuracy of item parameter estimates. These results suggested that reinstating the MAR assumption is important when calibrating items in MST.

Key Limitations: It remains unclear why the multiple-group Expectation Maximization (EM) algorithm fails to achieve acceptable parameter recovery, highlighting the need for effective methods to calibrate item parameters per subscale. Routing based on true θ in Design I was unrealistic in practice, although it illustrates conditions in which a multiple-group approach may be theoretically necessary under a MNAR scenario. Moreover, the study employed only one form per module in the MST design, whereas multiple parallel forms per module are common in practice. The simulations used the 2PL model throughout, limiting the generalizability of the results to other IRT models. Additionally, the number of items per subscale per block was too small to adequately recover the underlying θ distribution for each group. Future research could explore multidimensional IRT (MIRT) calibration, which considers all item responses within a routing block simultaneously and may provide an alternative solution for subscale calibration.

Wang, C., Zheng, Y., Chang, H.-H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika*, *79*(1), 154-74.

Description: This study introduced a new index–the SD of test overlap rate–as an additional measure of test security. The study analytically derived lower bounds for both the mean and the SD of test overlap rate under MST and CAT designs as well as an upper bound for the organized item theft index in MST. Although the mean overlap rates are theoretically equivalent for MST and CAT, the SD of test overlap for MST is expected to be larger than that of CAT. To provide empirical evidences, the study conducted simulations. For MST, a 1-3-5 panel was assembled with 10 items per module. To achieve uniform item exposure rates, 15, 5, and 3 parallel forms were constructed for the first-, second-, and third-stage modules, respectively. Routing was performed using the match-b-criterion method, where subsequent modules were selected by aligning average module difficulty with the examinee's ability estimate. For CAT, subsequent items were selected adaptively using the continuous a-stratification index (Wang& Chang, 2008) and the maximum priority index (Cheng & Chang, 2009). For each condition, 1,000 examinees were generated from $N(0, 1)$. Results were evaluated based on the average number of overlapping items per each possible pair of examinees and the SD of the test overlap. Additional studies were also carried out to investigate the relationship between SD of the test overlap rate, organized item theft size, and item pool size.

Key Findings: Based on the simulation study results, the findings aligned with those derived analytically; the mean test overlap rates were identical for MST and CAT, but the SD of test overlap rates was larger for MST. The study also showed that the SD of test overlap rates increased in the presence of organized item theft, suggesting that a larger SD indicates greater item sharing among examinees. Moreover, within MST, employing a two-pool design (as opposed to a single-pool) produced higher means and SDs of test overlap rates, indicating that non-overlapping multiple pools may elevate test security risk.

Key Limitations: The analytical results were derived under two assumptions: (1) each module within a stage had an equal probability of selection and (2) all forms of a module were strictly parallel. Also, for simulation, the module-length was consistent across stages.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Unpublished doctoral dissertation). Michigan State University.

Description: This study compared the performance of CAT and MST when item pools were designed separately to meet content specifications, CSEM, exposure rate constraints, and IRT scoring method. For the MST simulations, the study varied test length (45 vs. 60 items), panel design (1-2-3 vs. 1-3-3), routing strategy, and assembly priority. For routing strategy, the study considered the AMI and DPI methods. For assembly priority, the study implemented forward assembly (earlier to later stages) and backward assembly (later to earlier stages) using the bottom-up strategy. For CAT simulation, subsequent items were selected using the maximum priority index (MPT) method. Tests were terminated once an examinee's conditional test information fell within the 5% of the corresponding conditional test information under MST. The maximum test lengths for CAT were 55 and 70 items, corresponding to the 45- and 60-item MST conditions, respectively. For each study condition, 5,000 simulees were generated from $N(0, 1)$ with ten replications; and abilities were estimated via MLE. Results were evaluated in terms of ability estimation accuracy and average test length.

Key Findings: In terms of measurement accuracy, MST performed comparably to CAT. However, the study results indicated that CAT could achieve similar levels of conditional test information with a shorter average test length; the pattern was more evident as test length increased. Additionally, the findings suggested that the backward assembly method should be avoided when constructing MST panels, even for classification test.

Key Limitations: The study examined only a limited set of factors and conditions in assembling MST panels. All content areas and modules contained the same number of items. Moreover, both MST and CAT item pools and panels were constructed under the assumption of having a sufficient number of items–an assumption that may not hold in operational testing. Furthermore, the study focused exclusively on MC items.

Wang, Q. (2019). *The effects of test speedness control within a computerized adaptive multi-stage framework* (Unpublished doctoral dissertation). The University of Minnesota.

Description: The study proposed a new MST panel design aimed at controlling test speededness and compared its performance with that of traditional designs. For simulation, the study varied module length (10 vs. 20 items), panel design (1-3, 1-2-3, and 1-3-3), average difficulty of item pool (-1, 0, and 1), correlation between difficulty and time intensity (0 vs, .33), and speeded response pattern. For the speeded response pattern, the study considered three patterns: (1) moderate speedness, (2) resurgence symptom, and (3) adverse effect in the subsequent module. For each MST design, ten parallel panels were assembled using the MIP method to meet target TIF. Both interim and final abilities were estimated using the EAP method, and routing cutscores were determined using the AMI method. For comparison, a linear test form of equal length was also assembled for each panel design. The study generated 500 simulees for the linear forms and 2,000 simulees for the MST forms, all sampled from $N(0, 1)$. Results were evaluated in terms of ability estimation accuracy (bias, RMSE, and MAE) and item pool usage indices (percentage of item usage and item exposure rates).
Key Findings: Compared with the traditional MST designs, the speededness-controlled designs improved measurement precision but resulted in higher item pool exposure rates; however, this difference in the item pool exposure rate was minimal when pools were of ideal sizes. As the mean item difficulty increased from -1 to 1, the increases in measurement precision were greater for the proposed designs than for the traditional ones. In addition, ability estimation accuracy improved as the number of items increased.
Key Limitations: The proposed MST design could not accommodate multiple speeded response patterns; consequently, a single pattern was applied through the administration. Future research should explore varying degrees of test speededness and investigate methods to account for examinees' psychological adaptability in MST design (e.g., on-th-fly MST).

Wang, S., Lin, H., Chang, H.-H., Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45-62.

Description: The study proposed a hybrid adaptive framework by combining CAT and MST. In this framework, an initial MST step is followed by a CAT step, providing a more robust early testing stage than standard CAT and greater adaptivity in later stages than MST. The study presented two hybrid methods: (1) a preassembled MST followed by a CAT (PMCAT), and (2) a transition from an OMST (Zheng & Chang, 2015) to a CAT (FMCAT). Three simulation studies were conducted to evaluate the efficiency of the proposed methods. Study 1 and Study 2 assessed the efficiency of PMCAT and FMCAT, respectively, with a standard CAT design as the baseline. For the MST steps, the study used a 1-4 design; and the length for the MST step was varied to investigate its impact on ability estimation accuracy and efficiency. Study 3 compared the two hybrid designs with single designs, including a standard CAT, a CAT with Kullback–Leibler item selection method, a 1-4 MST, and a 1-2-4 MST. Estimation accuracy and efficiency were evaluated using RMSE, bias, and correct classification rate.

Key Findings: For the hybrid designs, estimation accuracy and efficiency were comparable to–or some cases exceeded– those of the traditional CAT and MST designs, particularly for examinees with extreme ability levels.

Key Limitations: The current study did not consider item exposure rates. In addition, the hybrid designs allowed examinees to review and change answers only during the MST step, not during the CAT step.

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

Description: This study evaluated the performance of multidimensional MST designs in terms of ability estimation accuracy and compared them with their unidimensional counterparts. A simulation study was conducted by manipulating four factors: design configuration (1-3, 1-2-3, and 1-3-3), number of items per dimension (3, 5, and 8), correlation among traits (0.2, 0.5, and 0.8), and item pool characteristics (μ_a = 0.6, 1 and μ_b = -1, 0). Throughout simulation, the study employed a simple structure measuring four latent traits. Based on the study factors, nine multidimensional MST designs were created, varying in design configuration and items per dimension. For each design, four item pools were generated by varying the means of discrimination and difficulty parameters, and ten parallel panels were assembled from each item pool. Across the 144 study conditions, the study simulated 3,000 examinees, with four latent traits drawn from a multivariate normal distribution. Each examinee was randomly assigned one of the ten parallel panels and its unidimensional counterparts for the four traits. Scoring under unidimensional designs employed both the 3PL UIRT and 3PL MIRT models, whereas the multidimensional designs used the 3PL MIRT model. Abilities were estimated using the EAP method. Results were evaluated in terms of ability estimation accuracy (bias and RMSE).

Key Findings: When correlations among traits were moderate to high (> 0.5), multidimensional MST designs provided greater precision in scoring than their unidimensional counterparts using UIRT scoring. However, their performance was comparable to their unidimensional counterpart with MIRT scoring. Based on the study results, the accuracy and efficiency of multidimensional MST designs improved as the item pool became more informative, suggesting that an optimal pool should include items that are both discriminative and appropriately difficult. The effect of test length (i.e., number of items per dimension) was not as large as that of trait correlation and item pool characteristics. Among the three MST configurations considered in this study, the 1-2-3 design seemed most promising.

Key Limitations: Given the limited factors and conditions examined in this study, the results should not be overgeneralized to other contexts.

Wang, Z., Li, Y., Wothke, W. (2019). Heuristic assembly of a classification multistage test with testlets. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 119-128). Springer Nature Switzerland AG: Springer.

Description: The study applied the modified NWADH and developed a heuristic procedure for assembling testlet-based multistage panels. The item bank consisted of items from operational tests that were calibrated using the bi-factor testlet (BFT) model. The study considered a 1-3-5 panel design which could classify examinees into six levels. Each module consisted of testlets totaling 10 items. The MST panel was constructed using the backward assembly method. To balance item bank usage, five parallel forms were constructed for first-stage modules and two for second-stage modules. Abilities were estimated using the BFT model and classified into one for the six levels at the end of testing. Routing cutscores were determined based on the intersections of information functions for adjacent follow-up modules. The study also included four additional tests: 44- and 30-item linear tests with prior information about examinee levels, and 44- and 30-item randomly selected tests. For the tests with prior information, hard and easy versions were constructed. For each study condition, 500 simulees were generated from $N(0, 1)$ and completed the five tests. Results were evaluated in terms of ability estimation accuracy and classification accuracy.

Key Findings: Based on the study results, the 30-item MST performed similarly to the 44-item linear test with prior information and outperformed the other linear tests. For classification purposes, MST allowed the test to be shortened without compromising outcomes.

Key Limitations: The item bank did not have enough items providing adequate information at the anchor points for the first- and second-stage modules. As a result, both the linear test with prior information and MST had to administer roughly the same items to high-ability examinees. Additionally, using the 2PL testlet model could provide accurate ability estimates for the 30-item test, suggesting that increasing the test to 40 items would offer little improvement.

Waters, B. K. (1975). *Empirical investigation of the stradaptive testing model for the measurement of human ability* (Unpublished doctoral dissertation). Florida State University.

Description: This study investigated the utility and validity of the stratified adaptive computerized testing model (stradaptive), a tailored testing strategy. The item pool was constructed based on nationally normed School and College Ability Test Verbal analogy items (SCAT-V). A total of 103 freshmen subjects were randomly assigned to one of two groups: conventional test group and stradaptive test group. Participants in the conventional group completed a fixed SCAT-V test, while those in the stradaptive group received individually tailored tests. All items for both test types were drawn from the same item pool. To compare the performance of the two groups (i.e., two tests), results were evaluated using KR-20 and parallel-form reliability coefficients.
Key Findings: Based on the study results, the reliabilities for the stradaptive group were significantly higher than those for the conventional group. Moreover, while administering fewer items, the stradaptive model achieved validity indices comparable to the conventional tests. However, participants in the stradaptive group required significantly more time to complete the items than those in the conventional group.
Key Limitations: It was recommended to conduct future research comparing termination rules for variable-length and fixed-length tests.

Weber, P. A. R. (2009). *Content clustering of a computerized-adaptive test* (Unpublished doctoral dissertation). Bethel University.

Description: This study examined whether a content-clustered computerized adaptive test, Measures of Academic Progress (MAP), would enhance performance among young test takers and improve testing efficiency. Test-taker performance was compared across demographic subgroups, including as gender (boys and girls), ethnicity (Caucasian, African American, and Hispanic), and grade level (second to fifth grades). Both the standard and modified versions of the MAP test were delivered to two halves of 6,589 students in second through fifth graders from a large school district in the southern United States.

Key Findings: The results indicated that content clustering did not improve overall MAP scores or provide significant time gains. However, there existed subgroup differences: African Americans performed better on the modified version; and, boys also showed score improvements with the modified version.

Key Limitations: The current study focused solely on the math portion of the standard MAP tool; therefore, the results should not be generalized to other content areas or to non-CAT assessments. Furthermore, the study could relied only on data from districts currently using MAP and those that had purchased the math assessment for grades two through five. Consequently, the findings are reliable only for the included sample and may not generalize to the entire population.

Weissman, A., Belov, D. I., Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests* (LSAC Research Report No. 07-05). Law School Admission Council.

Description: This study introduced a new routing method of maximizing mutual information, and compared the performance of three routing methods in terms of correct and incorrect classification rates, as well as the distribution of examinees across proficiency categories. For routing examinees, the study considered three methods utilizing: (1) NC, (2) MFI, and (3) maximum mutual information (MMI). To define ability points with maximum information, the MFI method uses a point estimate of an examinee ability, whereas the MMI method evaluates information at each of the points used in the likelihood ratio tests adopted for classifying examinees. For the simulation, the study used a 1-2-2-3 MST panel with four possible paths. Each module contained a testlet of 10-14 items; and each examinee was administered to a total of 44 to 50 items. The study simulated 5,000 examinees from $N(0, 1)$, and they were classified into one of three proficiency categories (lowest, middle, and highest) based on Neyman-Pearson likelihood ratio tests at the end of testing.

Key Findings: The study results showed that classification rates were similar across all three routing methods. Therefore, the study recommended using the NC routing method in practice due to its ability to control item exposure.

Key Limitations: The authors did not explicitly mention the limitations of their study.

Wu, I.-L. (2001). A new computer algorithm for simultaneous test construction of two-stage and multistage testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 180-198.

Description: This study proposed a new approach for simultaneously generating MST test forms from an item bank. The study explored the approach based on two binary programming models with a special network structure. In Model 1, TIF was maximized at a single specified ability point, whereas in Model 2, the target TIF (TTIF) was matched as closely as possible across multiple specified ability points. Both models were solved using an efficient special-purpose network algorithm. The item bank consisted of 520 items from 13 previously administered ACT math tests. To evaluate Model 1, the study conducted three experiments, each generating five 40-item tests with progressively increasing difficulty at rates of 0.3, 0.4, and 0.5. Each test was required to represent six taxonomy groups with 8, 14, 4, 8, 4, and 2 items, respectively. The five tests were constructed simultaneously using Model 1. For Model 2, additional ability points were incorporated across the 15 tests (3 experiments $\times$ 5 tests). Model 1 performance was assessed based on the achieved TIF at the specified ability point, while Model 2 performance was evaluated by comparing the specified TTIF with the achieved TTIF at multiple ability points.

Key Findings: The results indicated that Model 2 was less influenced by the properties of different item banks compared with Model 1. Overall, the new simultaneous approach for generating MST test forms seemed to outperform the traditional method in terms of test quality and computational efficiency.

Key Limitations: The results from Model I were dependent on the properties of the item bank, suggesting that applying it to different banks could yield varying outcomes. Although Model 2 was generally more stable, its results could still vary across different item banks. Future research could extend Model 1 and Model 2 to incorporate additional practical test constraints, such as maximizing or matching subtest information functions for each taxonomic group within a test.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Description: This study compared the performance of different CBT designs in the context of credentialing tests. For simulation, the study manipulated item quality level, item bank size, and passing scores. For item quality, the study considered three levels of the average a-parameter: 0.81, 1.06, and 1.31. For item bank, two sizes were considered: 240 and 480 items. Passing scores were set at three θ points (0, 0.5, and 1) corresponding to pass rates of 50%, 70%, and 85% for examinees drawn from $N(0, 1)$. For CBT designs, the study considered LPFT, MST, and CAT designs. For LPFT, five parallel forms of 35 items were created to have an item exposure rate of 0.2. For MST, the study considered 1-2 and 1-3 designs and assigned 20 and 15 items to the first- and second-stage modules, respectively. Routing cutscores were selected to assign roughly equal proportions of examinees to the second-stage modules. Multiple parallel modules were created to control item exposure rates. For CAT, the test length was also fixed at 35 items, with subsequent items selected using the maximum information procedure. For each crossed condition, 3,000 examinees were simulated from $N(0, 1)$ and abilities were estimated using the EAP method. Results were evaluated in terms of decision accuracy (DA), decision consistency (DC), and item bank utilization.
Key Findings: The study results were comparable and satisfactory across all the three designs. For each design, both DC and DA increased as item quality improved. When the item bank size increased, item exposure rates were significantly decreased; however, the improvements in DC and DA were minimal. As passing scores increased from 0 to 1, both DA and DC improved. The study also indicated that DA tended to be more affected by the target information function than by the item bank quality.
Key Limitations: The study conducted the simulation solely in the context of credentialing exams and did not evaluate performance using other criteria, such as the accuracy of ability estimates.

Xing, D., Hambleton, R. K. (2004). Impack of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*(1), 5-21.

Description: This study compared the performance of different CBT designs in the context of certification testing. For simulation, the study varied item bank size (240 vs. 460 items), item quality (average discrimination parameters of 0.60, 1.00, and 1.40), and CBT designs. As a result, there were six item banks, all calibrated using the 3PL IRT model. The CBT designs included LPFT, MST, and CAT. For LPFT, the study constructed five non-overlapping parallel forms of 35 items. For MST, a 1-3 panel was constructed with 20 and 15 items in the routing and second-stage modules, respectively. To control item exposure rate in the routing module, five parallel routing forms were built and randomly administered to examinees. Routing cutscores were determined so that three second-stage modules were administered to approximately equal numbers of examinees. For CAT, the test length was also fixed at 35 items. For each crossed study condition, 3,000 examinees were randomly drawn from $N(0, 1)$ and administered all three CBT designs. Scoring was conducted using the EAP method, and the passing score was set to achieve a 50% passing rate for the population. Results were evaluated in terms of decision accuracy (DA) and decision consistency (DC).

Key Findings: The results indicated that item bank size and quality had greater impact on DA and DC than the choice of test design. As item bank size and/or quality increased, both DA and DC improved across all three designs. The study results also suggested that CAT or MST offered no clear advantage over LPFT as long as the target measurement precision was achieved at the passing score. The authors concluded that investing in expanding the size and quality of item banks would be crucial for improving credentialing exam outcomes.

Key Limitations: The best bank with an average discrimination parameter of 1.40 could be unrealistic in practice, suggesting that some ideal conditions might not full reflect real-world testing scenarios. Additionally, the findings were limited to the context of credentialing exams, where the focus is solely on pass-fail decisions. Future research should explore other applications, such as diagnostic score reporting, awarding prizes, and reporting individual scores alongside pass-fail decisions.

Xiong, X. (2018). A hybrid strategy to construct multistage adaptive tests. *Applied Psychological Measurement*, *42*(8), 630-643.

Description: The study proposed a new hybrid strategy for assembling optimal MST panels, classifying test form requirements into six levels–solution, panel, route, module, item-set, and item levels–and simultaneously considering all requirements during panel assembly. The study applied the new strategy to a large-scale licensure examination consisting of four sections. Sections A, B, and C included multiple-choice questions (MCQs) and task-based simulations (TBSs), while Section D included MCQs and constructed-response (CR) items. For each section, the study applied both the hybrid strategy and the traditional bottom-up approach, and generated 24 parallel panels using operational item pools. The bottom-up approach was implemented using IBM ILOG CPLEX Optimization Studio 12.6.1, and the MIP technique was used for both methods. Results were evaluated in terms of TIFs and item pool utilization.
Key Findings: The hybrid method maintained test specifications while constructing optimal MST panels efficiently. Additionally, it enhanced item pool utilization, achieving higher selection rates for previously inactive items.
Key Limitations: Since the hybrid method considers all requirements simultaneously, the optimization process takes longer than the bottom-up approach.

Xu, T. (2010). *A review of exposure control strategies for CAT and potential applications in MST* (AICAP Technical Report). American Institute of CPAs.

Description: The study extended exposure control strategies for CAT and proposed two designs that could be applied for MST. Design 1 employed the a-stratified method with content blocking (Yi & Chang, 2003) to maintain content balance and control item exposure rates. A pilot study was conducted using a master pool from a CPA exam to demonstrate the application of Design 1. Design 2 divided a master pool into multiple parallel sub-pools and suggested the use of a single sub-pool for assembling test forms for each administration.

Key Findings: Design 1 raised concerns about examinee fatigue, potential routing errors, and challenge of deriving feasible target TIF values for each module. For Design 2, rotating item pools could be time consuming.

Key Limitations: The study was preliminary, and future research should examine the practical application of the proposed methods for MST.

Yamamoto, K., Khorramdel, L., Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347-368.

Description: To illustrate an MST adaptive design for international large-scale assessments (ILSAs), the study focused on the PIAAC (Programme for the International Assessment of Adult Competencies) and demonstrated how adaptive features could be integrated with the general requirements and restrictions of PIAAC (e.g., domains and adaptiveness within each domain). For instance, in MST designs for PIAAC, adaptiveness should be considered for the Literacy and Numeracy domains, but not for the domain of problem solving in technology-rich environment. Procedures for assembling MST forms for PIAAC were quite complicated (please see the original article for details). The current study obtained data from 24 countries and calibrated dichotomously- and polytomously-scored items using the 2PL model and GPCM, respectively. Performance of MST was evaluated in terms of its efficiency relative to linear tests and item position effects. Furthermore, to ensue score comparability across countries and languages, the study also examined item-by-country/language interactions.
Key Findings: The MST design for PIAAC improved measurement precision, particularly for low- and high-performing students and countries. The study found small item position effects. The study also demonstrated high comparability of item parameter estimates across countries and languages, suggesting that test scores were comparable across different groups.
Key Limitations: Since the MST design for PIAAC accounted for constraints typical for international large-scale assessments, its adaptiveness was somewhat limited. Future research should explore the use of larger item pools and refine adaptive procedures to enhance measurement precision across a broader range of examinee abilities.

Yamamoto, K., Shin, H. J., Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16-27.

Description: This article presented advantages of MST designs over CAT and fixed-test designs in the context of international large-scale assessments (ILSAs). The study illustrated unique features of MST designs for the PIAAC and the Programme for International Student Assessment (PISA), and discussed expected gains of using MST designs in ILSAs. Please refer to the article for a detailed description about MST designs considered in the study. Furthermore, it addressed limitations and challenges pertaining to ILSAs due to characteristics in its data collection design: the tests should be administered every three years in more than 80 countries through a type of balanced incomplete block (BIB) design. To evaluate results, the study calculated the bias of model parameters (i.e., mean and SD for each group as well as item slopes and difficulty parameters) and quantified the expected gains in ability precision by calculating the proportion of the MST design's standard error against a baseline design.

Key Findings: For PIAAC, the MST design was found to be 10–30% more efficient for Literacy and 4–31% more efficient for Numeracy than linear tests of equal length, with high comparability (92–97%) across countries. PISA 2018 simulations showed an expected average precision gain of about 4.5% for the person ability estimator across all scale scores, with around 10% higher accuracy at extreme performance levels (below 300 and above 700 PISA scale scores). This is crucial for improving proficiency estimation of both extremely high- and low-performing students and countries

Key Limitations: Since items are not administered to randomly equivalent groups, standard item statistics are not comparable across items, within a country, or across countries without accounting for the differential proportions of the routing paths. This requires calculating proportion correct by standardizing adaptive path proportions and using a modified multiple-group IRT model sensitive to the adaptive path for accurate item parameter estimation. Furthermore, MST designs are not as optimal in efficiency and accuracy as item-level CAT if items are truly independent. However, MST was chosen in PISA because it is difficult to break apart units comprising multiple items. The accuracy of routing procedures is also limited; for instance, human-coded constructed response items were not considered in PISA's routing decisions. Future study can consider optimizing threshold values that classify respondents into different groups and the proportion of randomization through simulation studies. The complexity of multilevel variables (e.g., country, language, education) in ILSAs also poses a risk of uneven exposure of test booklets and instability of scaling if the assessment is purely adaptive, thus requiring a balance between adaptiveness and full coverage of subpopulation abilities.

Yan, D. (2010). *Investigation of optimal design and scoring for adaptive multiii-stage testing: A tree-based regression approach* (Unpublished doctoral dissertation). Fordham University.

Description: Building on the tree-based CAT, this study proposed a nonparametric tree-based algorithm for routing examinees in the context of MST. For a three-stage 1-2-3 MST design, the study varied module lengths (10-15-20, 15-15-15, and 20-15-10) and ranges of module difficulties at each stage (wider and narrower). For each MST configuration, 250 examinees were randomly selected from a large operational assessment dataset and, cutscores between stages were determined using the tree-based algorithm. Total scores were then estimated using multiple linear regressions on observed NC scores across all modules for each subsample. The algorithm's performance was evaluated using the RMSE between observed and predicted scores.

Key Findings: Compared with a fixed 100-item linear test, the tree-based MST produced reliable scores using fewer items. These results suggest that the new approach is feasible for MST even without the unidimensionality and model assumptions of item response theory.

Key Limitations: The items used in the study were moderately discriminating; most of them had bi-serial correlations between .4 and .6. The items were also generally easy. The study did not address other practical considerations, such as content balance or item exposure rates. Future research should explore MST panel designs beyond the 1-2-3 configuration and account for population variation.

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* (Unpublished doctoral dissertation). Michigan State University.

Description: This study applied the p-optimality method (Reckase, 2003, 2010) to design MST optimal item pools and compared the performance of different MST designs. For simulation, the study varied MST panel design (1-2, 1-3, 1-2-2, 1-2-3), test length (20, 40, and 60 items), routing test proportion (20%, 30%, and 40%), and the presence or absence of module-level item exposure control (20% vs. no control). An additional empirical study was conducted for the 1-2-2 design with 75 items (25 items per module) in the context of a licensure exam. For both the simulation and empirical studies, the p-optimality method was applied to generate optimal item pools for each panel design. Abilities were estimated using the MLE method, and routing cutscores were set so that examinees were equally distributed across subsequent modules. Panels were assembled using a bottom-up approach that considered both statistical and non-statistical requirements. A total of 5,000 simulees were generated from $N(0, 1)$ and administered the tests. Results were evaluated in terms of both overall and conditional statistics. Overall statistics included correlation between θ and $\hat{\theta}$, test information, SEM, marginal reliability, item overlap and exposure rates, bias and RMSE in ability estimates, and classification accuracy. Conditional statistics included bias and RMSE in ability estimates, classification accuracy, SEM, and item overlap rate.

Key Findings: The p-optimality method proved effective for constructing optimal item pools, enabling the development of parallel modules and panels while maintaining sufficient measurement accuracy in ability estimation and classification. The method was also demonstrated feasibility for item pool design in operational testing contexts, such as licensure exams. With or without item exposure control, the maximum item exposure rate was 0.20; however, applying exposure control required item pools that were 7 to 11 times larger than those without control. Across study conditions, correlation between θ and $\hat{\theta}$ was consistently high. Both overall and conditional statistics were not affected much by MST designs, routing test proportions, and the presence of item exposure control. However, longer test lengths led to higher reliability and improved accuracy in ability estimation and classification.

Key Limitations: The generalizability of the findings should be limited within the scope of this study. The study did not consider content balancing in the design of the optimal item pools, and the test lengths and routing test proportions examined were not exhaustive. Future research could explore not only additional fixed-lengths but also variable-length MST structures. Moreover, this study was limited to dichotomously scored items modeled under the Rasch framework. Extensions could investigate alternative item types, including other dichotomous IRT models, polytomously scored or

testlet-based items, and even mixed-format assessments.

Yang, L., Reckase, M. D. (2020). The optimal item pool design in multistage computerized adaptive tests with the p-optimality method. *Educational and Psychological Measurement*, *80*(5), 955-974.

Description: The study extended the p-optimality method to the MST contexts to develop optimal item pools across different panel designs. For simulation, the study varied MST panel structure (1-2, 1-3, 1-2-2, and 1-3-3), test length (20, 40, and 60 items), routing module proportion (20%, 30%, and 40%), and the presence or absence of item exposure control. For three-stage structures,the third-stage module length was fixed at 40%. For each design, the p-optimality method was applied to construct optimal item pools. Routing cutscores were et to evenly distribute examinees across subsequent modules. The performance of the optimal item pools were evaluated in terms of ability estimation precision, classification accuracy at three score points (median value, 80th percentile, and 95th percentile of the observed score distribution), and item exposure and overlap rates. For overall evaluation, 5,000 examinees were randomly generated from $N(0, 1)$; and for conditional evaluation, 100 examinees were placed at fixed θ points ranging from -3.5 to 3.5 in increments of .05.
Key Findings: The performance of the p-optimality method improved as the test length increased. However, accuracy in ability estimates and classification was only minimally impacted by panel structure and the length of the routing modules. The p-optimality method proved effective in constructing optimal item pools for MST. Moreover, the bin width could be adjusted easily to align with the desired level of measurement precision.
Key Limitations: The scope of this study was limited. Future research could extend the p-optimality method to other IRT models, both fixed- and variable-length MST designs, and alternative item formats (e.g., polytomously scored or testlet-based items). In addition, the study could consider incorporating content balancing into the construction of optimal item pools for MST using the p-optimality method.

Zeng, W. (2016). *Making test batteries adaptive by using multistage testing techniques* (Unpublished doctoral dissertation). The University of Wisconsin-Milwaukee.

Description: The study proposed two test battery designs that incorporated MST components: the multistage test battery (MSTB) design and the hybrid multistage test battery (MSTBH) design. In the MSTB design, the battery consisted of three 1-2-4 MST tests. Based on the final ability estimate from the first test, the second and third panels were constructed using the OMST. In the MSTBH design, a battery also consisted of three test, with the first two tests administered as MST and the third test as CAT. Each test contained 24 items. Three module-length configurations were considered for the first, second, and third stages: 12-6-6, 6-6-12, and 8-8-8. Modules were constructed using the NWADH method combined with bottom-up and top-down approaches to maximize information at anchor points associated with each module. Ten parallel panels were created for each MST design. Interim and final ability estimates for both MST and CAT were obtained using two estimation methods: EAP and MLE. For the simulation study, item parameters were drawn from a retired operational item pool. For each study condition, 0,000 examinees were generated from $N(0, 1)$ per replication, with ten replications. The performance of the MSTB and MSTBH designs were evaluated in terms of measurement accuracy (classification accuracy, conditional correct classification rate, and RMSE in ability estimates) and test security properties (mean item exposure rate, item usage rate, overall and conditional overlap rates). Results were compared against two baseline designs: an MST design without borrowing information from previous tests and a CAT borrowing from previous tests (CATB).

Key Findings: The study results indicated that the two proposed battery designs achieved higher measurement accuracies than the baseline designs. When collateral information was incorporated, ability estimates obtained using the EAP method were more stable than those from the MLE method. Mean item exposure rates were acceptable across all designs; however, optimal pool utilization was achieved only for the CATB design.

Key Limitations: The study did not examine item usage across designs. Future research could extend item selection algorithms from CAT to the MSTB designs and explore alternative strategies for controlling item exposure rates. In addition, the study focused only on stand-alone items; future work could consider other item types, such as testlest.

Zenisky, A. (2004). *Evaluating the effects of several multistage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Description: The study examined the performance of MST design variables in the context of credentialing exams. For MST designs, the study varied module arrangement, target TIF, stage-level information, and routing strategy. For module arrangement, the study used the 1-2-2, 1-2-3, 1-3-2, and 1-3-3 designs. For the amount of target TIF, the study considered four conditions: the average target TIF from six operational forms, as well as three variations with a 50% increase, a 25% decrease, and a 50% decrease relative to the average. Stage-level information was distributed either equally across stages or in a 1/2, 1/4, 1/4 allocation to Stages 1, 2, and 3, respectively. For routing strategies, the study included the DPI, proximity, NC scoring, and random assignment methods. Across all designs, test length was fixed at 60 items, with 20 items per stage. For each crossed study condition, 9,000 simulees were randomly generated from $N(0, 1)$. At the end of testing, final abilities were estimated using the MLE method, and pass/fail decisions were based on three passing rates of 30%, 40%, and 50%. Results were evaluated in terms of decision accuracy (DA) and consistency (DC), accuracy of ability estimates, and the distribution of examinees across paths for each routing approach.

Key Findings: The study results indicated that measurement results–DA, DC, and accuracy in ability estimates–improved as target TIFs increased. Forhigh target TIF levels, performance was better when information was equally distributed across stages. At lower target TIF levels, similar results were achieved when half of the target TIF was allocated to Stage 1. Among the four routing approaches, the random assignment approach produced slightly lower measurement and decision accuracies, though the difference was not substantial. The DPI approach performed marginally worse than the proximity and NC scoring methods, whose performances were nearly identical. The study recommended the proximity approach for making high-stake decisions. Module arrangement had minimal impact on outcomes. Across different passing rates, DA, DC, and ability estimate accuracy tended to improve as passing rates decreased, i.e., as the passing scores moved away from the center of the distribution.

Key Limitations: As this was a simulation study with a limited set of conditions, the findings should be interpreted with caution and not generalized to other contexts.

Zenisky, A., Hambleton, R. K. (2004). *Effects of selected multi-stage test design alternatives on credentialing examination outcomes.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Sandiego, CA.

Description: This study compared the performance of MST aacross different design variables with respect to pass-fail decision and ability estimate accuracy. For simulation, the study varied panel design (1-2-2, 1-2-3, 1-3-2, and 1-3-3), target TIF level (baseline, 50% increase, 25% decrease, and 50% decrease), stage-wise TIF distribution ($\frac{1}{3}$-$\frac{1}{3}$-$\frac{1}{3}$ and $\frac{1}{2}$-$\frac{1}{4}$-$\frac{1}{4}$ for Stage 1-Stage 2-Stage 3), and routing method (DPI, proximity, NC scoring, and random assignment). For each crossed study condition, 9,000 examinees were randomly drawn from $N(0, 1)$, and final abilities were estimated using the MLE method. Pass/fail decisions were based on three thresholds: 30%, 40%, and 50%. Results were evaluated in terms of ability estimation accuracy and decision accuracy.
Key Findings: The quality of pass-fail decisions was not affected much by the panel designs. Accuracy in ability estimates and decision accuracy declined as the overall amount of TIF decreased. For lower TIF levels, allocating more information to the first stage improved both ability estimate accuracy and pass-fail decision quality. Among the four routing methods, the proximity and NC approaches performed similarly and generally outperformed the other methods, while the random approach produced the least accurate results. Across different passing thresholds, results improved as passing scores moved away from the center of the ability distribution (i.e., lower passing rates).
Key Limitations: Since simulation was conducted with a limited set of study conditions, the findings should be interpreted within the scope of this study. Future research could explore additional stage-wise TIF distributions and examine interactions between population ability distributions and passing scores locations. Moreover, future studies could consider using multidimensional or polytomously scored items.

Zhang, Y. (2006). *Impacts of multidimensionality and content misclassification on ability estimation in computerized adaptive sequential testing (CAST)* (Unpublished doctoral dissertation). The University of Delaware.

Description: This dissertation examined the robustness of CAST when tests were constructed, administered, and scored under a unidimensional IRT model, while underlying item responses were actually multidimensional. The study used the 3PL IRT model with two compensatory abilities. For simulated data, multidimensionality was manipulated by varying the angle between two item clusters (30, 60, and 90), the number of items in each cluster (500-500, 700-300, and 900-100 for Cluster 1-Cluster 2), and the correlation between the two abilities (.3 and .7). First, the study created an item pool of unidimensional parameters using 1,000 items from one subject of the P&P Uniform CPA exam. For each angle condition, a corresponding two-dimensional item pool was constructed based on the relationship between unidimensional and two-dimensional item parameters. The ATA procedure was used to construct 1-2-2 CAST panels by manipulating the number of items per cluster; the 500-500 condition represented the true content specification of the item pool. For each module, the study used the NWADH method to select 20 items achieving the target TIF. Both unidimensional and multidimensional datasets were simulated. For unidimensional data, 5,000 abilities were sampled from $N(0, 1)$. For multidimensional data, abilities were generated from a bivariate normal distribution with the specified correlation, and responses were generated using a two-dimensional MIRT model. Examinees abilities were estimated assuming unidimensional item parameters. Performance was evaluated in terms of ability estimates, routing decision, and pass/fail decision.

Key Findings: Regarding pass/fail decision, a perfect match in content specification produced the lowest type I errors across all levels of multidimensionality. For ability estimates and routing decisions, the impact of content misclassification level was minimal under mild multidimensionality. However, with severe multidimensionality, panels without content misclassification generally yielded more accurate ability estimates and routing decisions. Overall, content matching had a greater effect on pass/fail decision accuracy, whereas its impact on unidimensional ability estimation and routing decisions was relatively minor.

Key Limitations: The generalizability of the findings should be limited within the scope of the study. Moreover, the results may not hold when multiple panels are drawn from an item pool. In addition, the study did not provide a clear quantification of multidimensionality, which was manipulated through several factors; dimensionality could be assessed on item response data using methods such as DIMTEST or DETECT.

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*, *78*(1), 37-58.

Description: This paper proposed a modified DETECT index for conducting dimensionality analyses on response data from test designs, including CAT and MST. A simulation study using a two-stage test was conducted to show that the modified DETECT could successfully recover the dimensional structure of response data under reasonable specifications. For simulation, the study considered a 1-3 design and varied sample size (750, 1500, 3000, and 4000) and number of dimensions (1, 2, and 3). For unidimensional data, simulees were generated from $N(0, 1)$; for multidimensional data, simulees were drawn from a multivariate normal distribution with mean 0, variance 1, and correlation .8. NC cutscores were set to route equal proportion of examinees to each of the three modules at the second stage. Additionally, the modified DETECT procedure was applied to real operational two-stage test data sample. Items were calibrated using the 2PL IRT model in PARSCALE, and PolyDETECT was employed to determine the number of dimensions.

Key Findings: The study results indicated that the modified DETECT could successfully handle response data from various test designs. The modified procedure successfully partitioned items according to their dimensionality. Moreover, for multidimensional response data, the modified DETECT index could be decomposed into two parts that could serve as reliability indices for the DETECT results.

Key Limitations: The proposed index assumed an approximate simple structure for items, allowing each item to measure multiple domains. Although Zhang and Stout (1999) demonstrated that DETECT remains applicable under moderate violations of this assumption, further research is needed to determine the extent of violation that can be tolerated while still yielding valid results.

Zhang, Y., Breithaupt, K., Tessema, A., Chuah, D. (2006). *Empirical vs. expected IRT-based reliability estimation in computerized multistage testing (MST).* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Description: This study compared two IRT-based procedures for estimating test reliability in a certification exam that employed both non-adaptive and adaptive (via an MST model) designs. Both procedures relied on calibrated item parameters to estimate error variances. In terms of score variance, Method 1 used an empirical ability distribution of a specific examinee sample, whereas Method 2 was a sample-free procedure assuming a normal ability distribution. An extended version of Method 1 (modified Method 1) was also considered to address sampling restrictions in adaptive tests and to estimate reliability for each testlet within an MST panel before aggregating estimates into an overall reliability for a test form or a route. To illustrate the methods, the study used data from a high-stakes computerized certification exam that consisted of three adaptive sections (1-2-2 design), followed by a non-adaptive section.

Key Findings: The study results indicated that Method 1 and Method 2 produced similar results at both the panel and test-section levels for both adaptive and non-adaptive tests. However, Method 1 was not recommended for individual MST test forms; instead, the modified Method 1 is preferred, as it mitigates the impact of restricted samples.

Key Limitations: The authors did not provide a theoretical rationale for choosing between Method 1 or Method 2, as this was beyond the scope of this study.

Zheng, Y., Chan, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118.

Description: The authors proposed a new MST method, termed "on-the-fly assembled multistage adaptive testing" (OMST), designed to address certain limitations of CAT and MST while combining their advantages. For CAT and OMST, simulation was conducted manipulating three factors: (a) constrained item selection approach (shadow test vs. maximum priority index methods); (b) exposure control algorithm (Sympson–Hetter (SH) vs. multinomial Sympson–Hetter (MSH)); and (c) item bank stratification (stratified vs. non-stratified methods). With a stratified item bank, the study employed a underused sub-bank to select 15 items for both the first stage of OMST and CAT, while the remaining 30 items were selected from a well-used sub-bank. For MST, this study used a 1-3-3 design with a fixed test length of 45 items. parallel modules and panels were assembled using the NWADH method combined with a bottom-up approach. For each condition, 5,000 abilities were generated from $N(0, 1)$ truncated to the interval (-3.5, 3.5), and responses were generated using the 3PL IRT model. When responses were all correct or incorrect, ability estimates were obtained using EAP, and MLE otherwise. The study conducted fifty replications. Results were evaluated in terms of measurement accuracy, maximum exposure rate, constraint violations, item bank usage, and test overlap rates.

Key Findings: In term of measurement accuracy, OMST performed comparably to CAT and MST. The study also demonstrated that item bank stratification improved item bank usage without substantially reducing measurement accuracy. Regarding maximum item exposure rate, the SH method outperformed the MSH method by maintaining the maximum item exposure rates at desirable levels. Test security was higher for OMST and CAT compared to MST. Additionally, the study showed that violation-free OMST could be achieved when the constraint-controlled item selection method was combined with the item replacement step.

Key Limitations: The results highly depended on the quality of the item bank and the specific designs considered in the study. Thus, although the findings suggested that OMST performed well, they do not establish the overall superiority of any testing mode.

Zheng, Y., Nozawa, Y., Gao, X., Chang, H.-H. (2012). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes* (ACT Research Report Series No. 2012-6). Iowa City, IA: ACT.

Description: This study examined the effects of various factors on MST performance and compared these results with linear test form and CAT. For simulation, the study varied MST design (1-2-3 and 1-2-3-4), module length (equal-length, longer early stages, longer middle stages, and longer later stages), item overlap (none vs. overlap), routing strategy (θ-score vs. true-score), and assembly priority (backward vs. forward). For the θ-score strategy, cutscores were determined using the AMI method, and examinees' $\hat{\theta}$s were compared to these cutscores. For the true-score routing strategy, cutscores in terms of θ were converted to true scores via TCCs, and examinees' NC scores were compared to those thresholds. Test lengths were fixed at 21 items across all conditions. Parallel panels were assembled using the top-down procedure (Luecht & Nungester, 1998) based on the NWADH method (Luecht, 1998). For each crossed conditions, 5,000 examinees were simulated from $N(0, 1)$ truncated to (-3.5, 3.5). Final abilities were estimated using the MLE method and examinees were classified into one of five categories. For CAT, examinees received one of the 15 most informative items consecutively until the test length reached 21 items. The linear test form consisted of 30 items. Throughout the whole procedure, the study used the 3PL IRT model. Results were evaluated in terms of ability estimate accuracy, classification accuracy, and item bank usage.

Key Findings: The study results indicated that the number of stages, module-length assignment, and routing strategy had minimal impact on performance. Classification accuracy was generally higher for backward assembly than for forward assembly and tended to be higher under the item overlap condition compared to no overlap. However, item bank usage was better in the no-overlap condition than in the overlap condition. Compared to the 30-item linear test, the 21-item MST designs yielded higher measurement accuracy. Relative to the 21-item CAT, no-overlap MST designs achieved more efficient item bank usage, though with slightly lower measurement accuracy.

Key Limitations: The conditions examined in this study were limited. Future research should explore additional conditions, such as variable MST test lengths, alternative test assembly methods, and different item types.

Zwitser, R. J., Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*(1), 65-84.

Description: The study demonstrated how conditional maximum likelihood (CML) could be applied to make statistical inference within MST designs, focusing on parameter estimation and model fit evaluation. Although Glas (1988) argued that CML could not be used with MST, this paper showed that his conclusion was incorrect by illustrating that both MML and CML produced consistent estimates of difficulty parameters. The study also emphasized the importance of evaluating parameter estimation and model fit. It explained that simple measurement models may fit adaptive designs better than linear designs for two reasons: 1) a better match between item difficulty and examinee ability, reducing undesirable response behaviors (e.g., slipping and quessing), and 2) a larger number of items and parameters relative to the same number of observations in adaptive designs. Both simulation and real-data studies were conducted to illustrate the properties of CML inferences. The simulation consisted of three examples, all based on a 1-2 MST design with a routing cutscore of .5. Each test had 30 items, with 10 in the first-stage module and 20 items in the second-stage module. Abilities were randomly sampled from a mixture of two normal distributions: $N(-1.5, 0.5)$ with probability 2/3, and $N(1, 1)$ with probability 1/3. Results from the three examples were used to compare the performance of three methods—MST CML, ordinary CML, and ordinary MML—in terms of the accuracy of item parameter estimates. Furthermore, the simulation study included likelihood ratio tests to assess model-fit and displayed QQ-plots to evaluate item fit.

Key Findings: The authors demonstrated that the CML method was applicable to data from MST designs based on the Rasch model. Adaptive designs, including MST, led to better model fit compared to traditional linear tests, partly due to the potential avoidance of undesirable response behaviors like guessing and slipping, and increased parameter availability for the same number of observations. The findings suggested using the CML method when the population distribution might be misspecified. Additionally, the study highlighted the robustness of adaptive designs against undesirable response behaviors.

Key Limitations: CML estimators are generally less efficient than MML estimators; when MML assumptions are satisfied, MML may be preferable to CML. Future research should investigate optimal designs that maximize the efficiency of person parameter estimation. Additionally, the study did not account for the estimation error related to item parameters, and future studies could explore designs with more modules and stages.

References

AlGhamdi, H. M. (2018). *Assessment of multiple-form structure designs of multistage testing using IRT* (Unpublished doctoral dissertation). The University of Denver.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement*, *30*(3), 204-215.

Armstrong, R. D. (2002). *Routing rules for multiple-form structures* (ETS Research Report No. 02-08). New Town, PA: Law School Admission Council.

Armstrong, R. D. (2005). *A method to determine targets for multi-stage adaptive tests.*

Armstrong, R. D., & Edmonds, J. (2004). *A study of multiple stage adaptive test designs.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

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.

Beard, J. J. (2008). *An investigation of vertical scaling with item response theory using a multistage testing framework* (Unpublished doctoral dissertation). The University of Iowa.

Belov, D. I., & Armstrong, R. D. (2008). A monte carlo approach to the design, assembly, and evaluation of multistage adpative tests. *Applied Psychological Measurement*, *32*(2), 119-137.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019a). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement*, *56*(1), 121-146.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019b). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontier in Education*, *4:1*.

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing*, *5*(3), 319-330.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5-20.

Cai, L. (2018). *An investigation of item calibration approaches in multistage testing* (Unpublished doctoral dissertation). The University of Nebraska.

Cetin-Berber, D. D., Sari, H. I., & Huggins-Manley, A. C. (2019). Imputation methods to deal with missing responses in computerized adaptive multistage testing. *Educational and Psychological Measurement*, *79*(3), 495-511.

Chen, L.-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Chuah, S. C., 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.

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968a). An exploratory study of programmed tests. *Educational and Psychological Measurement*, *28*, 345-360.

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968b). Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, *5*(3), 183-187.

Colvin, K. F. (2014). *Effect of automatic item generation on ability estimates in a multistage test* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multistage framework* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

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 No. RR-11-26). Princeton, NJ: Educational Testing Service.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the mcat. *Applied Psychological Measurement*, *27*(5), 335-356.

Du, Y., Li, A., & Chang, H.-H. (2019). Utilizing response time in on-the-fly multistage adaptive testing. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 107-117). Springer Nature Switzerland AG: Springer.

Edmonds, J. J. (2004). *The evaluation of multiple stage adaptive test designs* (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey.

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education*, *25*(2), 118-141.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica: International Journal of Methodology and Experimental Psychology*, *32*(1), 107-132.

Gierl, M. J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multistage computer adaptive testing. *Educational Research and Evaluation*, *19*(2-3), 188-203.

Glas, C. (1988). The rasch model and multistage testing. *Journal of Educational Statistics*, *13*(1), 45-52.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education*, *19*(3), 221-239.

Han, K. T. (2020). Framework for developing multistage testing with intersectional

routing for short-length tests. *Applied Psychological Measurement*, *44*(2), 87-102.

Han, K. T., Dimitrov, D. M., & Al-Mashary, F. (2019). Developing multistage tests using D-scoring method. *Educational and Psychological Measurement*, *79*(5), 988-1008.

Han, K. T., & Guo, F. (2013). *An approach to assembling optimal multistage testing modules on the fly* (GMAC Research Report No. RR-13-01). Reston, VA: Graduate Management Admission Council.

Hembry, I. F. (2014). *Operational characteristics of mixed-format multistage tests using the 3PL testlet response theory model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Hendrickson, A. (2002). *Scaling of two-stage adaptive test configurations for achievement testing* (Unpublished doctoral dissertation). The University of Iowa.

Hendrickson, A. (2007). An ncme instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

Jewsbury, P. A., & van Rijin, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics*, *45*(4), 383-402.

Jiang, Y. (2019). Statistical considerations for subscore reportingin multistage testing. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 129-136). Springer Nature Switzerland AG: Springer.

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pools* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203-220.

Kaplan, M. (2016). *New item selection and test administration procedures for cognitive diagnosis computerized adaptive testing* (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey.

Karatoprak Ersen, R., & Lee, W. (2023). Pretest item calibration in computerized multistage adaptive testing. *Journal of Educational Measurement*, *60*(3), 379-401.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Unpublished doctoral dissertation). The University of Texas at Austin.

Kim, H., & Plake, B. S. (1993). *Monte carlo simulation comparison of two-stage testing and computerized adaptive testing.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Kim, J. (2010). *A comparison of computer-based classification testing approaches using mixed-format tests with the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Kim, J., Chung, H., Dodd, B. G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, *72*(4), 574-588.

Kim, J., Chung, H., Park, R., & Dodd, B. G. (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.

Kim, S., & Livingston, S. A. (2017). *Accuracy of a classical test theory–based procedure for estimating the reliability of a multistage test* (ETS Research Report No. RR–17-02). Princeton, NJ: Educational Testing Service.

Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (ETS Research Report No. RR–14-01). Princeton, NJ: Educational Testing Service.

Kim, S., & Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (ETS Research Report No. RR-16-22). Princeton, NJ: Educational Testing Service.

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

Kim, S., Moses, T., & Yoo, H. (2015b). *Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing* (ETS Research Report No. RR–15-11). Princeton, NJ: Educational Testing Service.

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, *14*.

Li, G., Cai, Y., Gao, X., Wang, D., & Tu, D. (2021). Automated test assembly for multistage testing with cognitive diagnosis. *Frontiers in Psychology*, *12:509844*.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*(1), 129-146.

Lord, F. M. (1969). *A theoretical study of two stage testing* (ETS Research Bulletin Series No. RB-69-95). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*(3), 147-151.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, *36*(3), 227-242.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test* (ETS Research Report No. RB-74-30). Princeton, NJ: Educational Testing Service.

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions* (Unpublished doctoral dissertation). The University of Maryland College Park.

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2003). *Exposure control using adaptive multi-stage item bundles.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Luecht, R. M., & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249.

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243-263.

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing*, *19*(3), 227-247.

Ma, Y. (2020). *Investigating hybrid test designs in passage-based adaptive tests* (Unpublished doctoral dissertation). The University of Iowa.

Macken-Ruiz, C. (2008). *A comparison of multistage and computerized adaptive tests based on the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27.

Martin-Fernandez, M., Ponsoda, V., Díaz, J., Shih, P.-C., & Revuelta, J. (2016). A multistage adaptive test of fluid intelligence. *Psicothema*, *28*(3), 346-352.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185-187.

Park, R. (2013). *The impact of statistical constraints on classification accuracy for multistage tests.* Durham, NC: American Institute of CPAs.

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* (Unpublished doctoral dissertation). The University of Texas at Austin.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. *Applied Psychological Measurement*, *38*(4), 268-280.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2017). The development of MST test

information for the prediction of test performances. *Educational and Psychological Measurement*, *77*(4), 570-586.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, *50*(4), 447-468.

Reese, L. M., & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing. law school admission council computerized testing report.* (LSAC Research Report No. LSAC-R-96-04). Princeton, NJ: Law School Admission Council.

Rome, L. (2017). *Evaluating item selection methods for adaptive tests with complex content constraints* (Unpublished doctoral dissertation). The University of Wisconsin-Milwaukee.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* (ETS Research Report No. RR-07-04). Princeton, NJ: Educational Testing Service.

Sadeghi, K., & Khonbi, Z. A. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language Testing in Asia*, *7*.

Sari, H. I., & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory and Practice*, *17*(5), 1759-1781.

Sari, H. I., & Raborn, A. (2018). What information works best?: A comparison of routing methods. *Applied Psychological Measurement*, *42*(6), 499-515.

Sari, H. I., Sari, H. Y., & Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406.

Schnipke, D. L., & Reese, L. M. (1997). *A comparison of testlet-based test designs for computerized adaptive testing.* A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education*, *19*(3), 257-260.

Svetina, D., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192-213.

Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* (Unpublished doctoral dissertation). The University of Illinois at Urbana-Champaign.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*,

44(2), 117-130.

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3-28.

Wang, C., Zheng, Y., & Chang, H.-H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika*, *79*(1), 154-74.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Unpublished doctoral dissertation). Michigan State University.

Wang, Q. (2019). *The effects of test speedness control within a computerized adaptive multi-stage framework* (Unpublished doctoral dissertation). The University of Minnesota.

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45-62.

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

Wang, Z., Li, Y., & Wothke, W. (2019). Heuristic assembly of a classification multistage test with testlets. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, p. 119-128). Springer Nature Switzerland AG: Springer.

Waters, B. K. (1975). *Empirical investigation of the stradaptive testing model for the measurement of human ability* (Unpublished doctoral dissertation). Florida State University.

Weber, P. A. R. (2009). *Content clustering of a computerized-adaptive test* (Unpublished doctoral dissertation). Bethel University.

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests* (LSAC Research Report No. 07-05). Law School Admission Council.

Wu, I.-L. (2001). A new computer algorithm for simultaneous test construction of two-stage and multistage testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 180-198.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Xing, D., & Hambleton, R. K. (2004). Impack of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*(1), 5-21.

Xiong, X. (2018). A hybrid strategy to construct multistage adaptive tests. *Applied*

Psychological Measurement, *42*(8), 630-643.

Xu, T. (2010). *A review of exposure control strategies for CAT and potential applications in MST* (AICAP Technical Report). American Institute of CPAs.

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347-368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16-27.

Yan, D. (2010). *Investigation of optimal design and scoring for adaptive multii-stage testing: A tree-based regression approach* (Unpublished doctoral dissertation). Fordham University.

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* (Unpublished doctoral dissertation). Michigan State University.

Yang, L., & Reckase, M. D. (2020). The optimal item pool design in multistage computerized adaptive tests with the *p*-optimality method. *Educational and Psychological Measurement*, *80*(5), 955-974.

Yi, Q., & Chang, H. H. (2003). *a*-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*, 359-378.

Zeng, W. (2016). *Making test batteries adaptive by using multistage testing techniques* (Unpublished doctoral dissertation). The University of Wisconsin-Milwaukee.

Zenisky, A. (2004). *Evaluating the effects of several multistage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Unpublished doctoral dissertation). The University of Massachusetts Amherst.

Zenisky, A., & Hambleton, R. K. (2004). *Effects of selected multi-stage test design alternatives on credentialing examination outcomes.* A paper presented at the Annual Meeting of the National Council on Measurement in Education, Sandiego, CA.

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*, *78*(1), 37-58.

Zhang, Y. (2006). *Impacts of multidimensionality and content misclassification on ability estimation in computerized adaptive sequential testing (CAST)* (Unpublished doctoral dissertation). The University of Delaware.

Zhang, Y., Breithaupt, K., Tessema, A., & Chuah, D. (2006). *Empirical vs. expected IRT-based reliability estimation in computerized multistage testing (MST).* A paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Zheng, Y., & Chan, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H.-H. (2012). *Multistage adaptive testing for*

a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes (ACT Research Report Series No. 2012-6). Iowa City, IA: ACT.

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*(1), 65-84.