

1-1-2004

Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment.

April L. Zenisky
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Zenisky, April L., "Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment." (2004). *Doctoral Dissertations 1896 - February 2014*. 5710.

https://scholarworks.umass.edu/dissertations_1/5710

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST



312066 0289 0053 8

EVALUATING THE EFFECTS OF SEVERAL MULTI-STAGE TESTING DESIGN
VARIABLES ON SELECTED PSYCHOMETRIC OUTCOMES FOR
CERTIFICATION AND LICENSURE ASSESSMENT

A Dissertation Presented

by

APRIL L. ZENISKY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

MAY 2004

Research and Evaluation Methods

© Copyright by April L. Zenisky 2004

All Rights Reserved

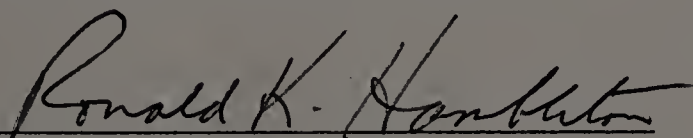
EVALUATING THE EFFECTS OF SEVERAL MULTI-STAGE TESTING DESIGN
VARIABLES ON SELECTED PSYCHOMETRIC OUTCOMES FOR
CERTIFICATION AND LICENSURE ASSESSMENT

A Dissertation Presented

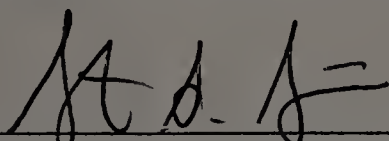
by

APRIL L. ZENISKY

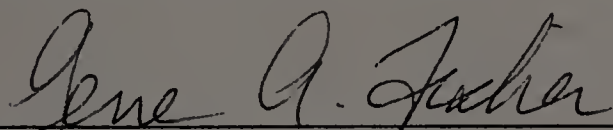
Approved as to style and content by:



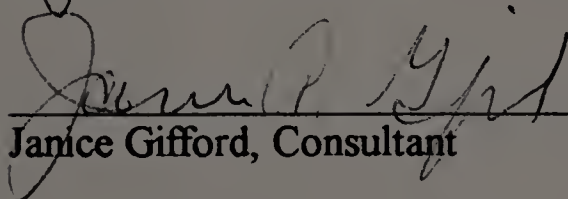
Ronald K. Hambleton, Chair



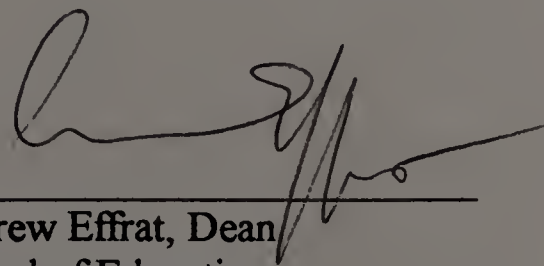
Stephen G. Sireci, Member



Gene A. Fisher, Member



Janice Gifford, Consultant



Andrew Effrat, Dean
School of Education

DEDICATION

For my parents, with gratitude.

ACKNOWLEDGMENTS

I begin these acknowledgements by expressing my deepest appreciation to the chair of my dissertation committee, my advisor throughout graduate school, and my professional mentor, Professor Ronald Hambleton. Both with respect to this study and the many other projects that Ron has involved me in, I am grateful for his ever-present willingness to share his time and considerable technical and practical knowledge with me to do things well. This study has benefited considerably from Ron's counsel and encouragement, and his support throughout has been invaluable to me.

As a member of my committee and a REMP professor, Stephen Sireci has always provided me with exceptional guidance and the great flexibility to finish this dissertation while working at the Center for Educational Assessment. I also thank Professors Gene Fisher and Janice Gifford for the significant insight they have contributed to this project.

I would be remiss not to acknowledge the generous financial and technical support given to this project by the American Institute of Certified Public Accountants, including Drs. Craig Mills, Jerry Melican, and Krista Breithaupt, as well as Dr. Richard Luecht through the AICPA-sponsored Research Consortium (for feedback on this research and use of CASTISEL). Professional Examination Service of New York, NY also provided support for the foundation research needed to complete this study.

The support and encouragement of a number of former and current students from REMP was essential to me in completing this study. I am especially grateful for my friendship with Lisa Keller, as I can always count on her for a smile and a lunch outside in the sunshine. Furthermore, I want to thank Lisa for her professional encouragement and willingness to share her technical expertise with me, especially in later stages when I

was mired in the blunter points of programming. Michael Jodoin's generosity with his time, knowledge, and software likewise helped me through many long days, as did his endless capacity for procrastinating with good conversation and even better beverages.

If there is one person whose actions have shown me the worth of pushing yourself to achieve, it is my younger brother, Matt. Through my admiration and respect for him, I have been encouraged in countless ways to go out and make good things happen.

Raldy, who keeps me sane when I start to make the little things big, has helped me in so many ways over the past several years to remember that this project is but one part of my life. From watching the sunrise on a Florida beach during spring training season to late nights with PlayStation hockey, he has given me many opportunities to be happy every day, and for that I am lucky.

Though there may be moments when we wonder whose kid I am, really, there can be no doubt: I am so proud to be the daughter of Charles and Nancy Zenisky. Together, their unwavering love and support for me in everything I have undertaken has given me the capacity and confidence to try so many things that I might never have otherwise seen myself doing. This degree, as with everything I have accomplished, would not have been possible without their presence, in every possible way. This is for you.

ABSTRACT

EVALUATING THE EFFECTS OF SEVERAL MULTI-STAGE TESTING DESIGN VARIABLES ON SELECTED PSYCHOMETRIC OUTCOMES FOR CERTIFICATION AND LICENSURE ASSESSMENT

MAY 2004

APRIL L. ZENISKY, B.A., AMHERST COLLEGE

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Computer-based testing is becoming popular with credentialing agencies because new test designs are possible and the evidence is clear that these new designs can increase the reliability and validity of candidate scores and pass/fail decisions. Research on MST to date suggests that the measurement quality of MST results is comparable to full-fledged computer-adaptive tests and improved over computerized fixed-form tests. MST's promise dwells in this potential for improved measurement with greater control than other adaptive approaches for constructing test forms.

Recommending use of the MST design and advising how best to set up the design, however, are two different things. The purpose of the current simulation study was to advance an established line of research on MST methodology by enhancing understanding of how several important design variables affect outcomes for high-stakes credentialing.

Modeling of the item bank, the candidate population, and the statistical characteristics of test items reflect an operational credentialing exam's conditions. Studied variables were module arrangement (4 designs), amount of overall test information (4 levels), distribution of information over stages (2 variations), strategies for between-stage routing (4 levels), and pass rates (3 levels), for 384 conditions total.

Results showed that high levels of decision accuracy (DA) and decision consistency (DC) were consistently observed, even when test information was reduced by as much as 25%. No differences due to the choice of module arrangement were found. With high overall test information, results were optimal when test information was divided equally among stages; with reduced test information gathering more test information at Stage 1 provided the best results.

Generalizing simulation study findings is always problematic. In practice, psychometric models never completely explain candidate performance, and with MST, there is always the potential psychological impact on candidates if test difficulty shifts are noticed. At the same time, two findings seem to stand out in this research: (1) with limited amounts of overall test information, it may be best to capitalize on available information with accurate branching decisions early, and (2) there may be little statistical advantage in exceeding test information much above 10 as gains in reliability and validity appear minimal.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
ABSTRACT	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiv
CHAPTER	
1. INTRODUCTION	1
1.1 Background	1
1.2 Statement of the Problem	8
1.3 Purpose of the Study	10
1.4 Significance of the Problem	11
2. REVIEW OF THE LITERATURE	15
2.1 Introduction	15
2.2 Fundamentals of the MST Design	15
2.2.1 Basic Structural Variables of the MST Design	19
2.2.2 Test and Module Assembly	22
2.2.3 Administration	30
2.2.4 Summary of MST Design Variables	34
2.3 Foundation Research in Adaptive-By-Stage Testing	34
2.4 Current MST Research	39
2.4.1 Evaluating MST Relative to Other Test Designs	41
2.4.2 Automated Test Assembly (ATA)	46
2.4.3 Special Applications of MST	48
2.4.3.1 Non-adaptive MST	48
2.4.3.2 Modules with Common-Stem Items	50
2.5 Summary	53
3. METHODOLOGY	57
3.1 Overview of the Study	57

3.2	Simulation of the Item Bank	61
3.3	MST Test Structures	63
3.4	ATA using CASTISEL	64
3.4.1	Obtaining the Base Target TIF	66
3.4.2	Targeting Test Information to the Passing Score.....	67
3.4.3	Amount of Target Test Information.....	68
3.4.4	Specifying Stage-Level Information.....	69
3.4.5	Quantifying Within-Stage Module Difficulty Differences	71
3.4.6	Content-Balancing Test Forms and Modules	72
3.4.7	Assembling Parallel Panels.....	73
3.4.8	Meeting Target TIFS.....	75
3.5	Simulating a Multi-Stage Test	76
3.5.1	MST Simulation with Varying Strategies for Between-Stage Routing....	77
3.5.1.1	Defined Population Intervals (DPI)	77
3.5.1.2	Matching Module Difficulty and Ability Estimates (Proximity).....	79
3.5.1.3	Number-Correct Scoring.....	80
3.5.1.4	Random Module Assignment	82
3.6	Computer Simulation Method.....	82
3.7	Data Analysis	84
3.7.1	Decision Accuracy	84
3.7.2	Decision Consistency	85
3.7.3	Accuracy of Ability Estimation	85
3.7.4	Simulee Routing Analysis.....	86
4.	RESULTS	94
4.1	Overview.....	94
4.2	Decision Accuracy	94
4.3	Decision Consistency.....	98
4.4	Accuracy of Ability Estimation	99
4.4.1	Correlations between True and Estimated Abilities	100
4.4.2	Root Mean Square Errors.....	101
4.5	Routing Path Analysis.....	104
4.5.1	Routing Path Analysis for the 1-2-2 Design Structure	104
4.5.2	Routing Path Analysis for the 1-3-3 Design Structure	105
4.5.3	Routing Path Analysis for the 1-2-3 Design Structure	106
4.5.4	Routing Path Analysis for the 1-3-2 Design Structure	106

4.6	Summary	106
5.	CONCLUSIONS.....	151
5.1	Conclusions.....	151
5.2	Directions for Future Research	156
	BIBLIOGRAPHY	158

LIST OF TABLES

Table	Page
3.1 Guidelines for Approximate Content Balancing Within and Across Stage	88
4.1 Decision Accuracy for DPI Routing at 30% Passing.....	109
4.2 Decision Accuracy for Proximity Routing at 30% Passing	110
4.3 Decision Accuracy for Number-Correct Routing at 30% Passing.....	111
4.4 Decision Accuracy for Random Routing at 30% Passing	112
4.5 Decision Accuracy for DPI Routing at 40% Passing.....	113
4.6 Decision Accuracy for Proximity Routing at 40% Passing	114
4.7 Decision Accuracy for Number-Correct Routing at 40% Passing.....	115
4.8 Decision Accuracy for Random Routing at 40% Passing	116
4.9 Decision Accuracy for DPI Routing at 50% Passing.....	117
4.10 Decision Accuracy for Proximity Routing at 50% Passing	118
4.11 Decision Accuracy for Number-Correct Routing at 50% Passing.....	119
4.12 Decision Accuracy for Random Routing at 50% Passing	120
4.13 Decision Consistency for DPI Routing at Three Pass Rates.....	121
4.14 Decision Consistency for Proximity Routing at Three Pass Rates	122
4.15 Decision Consistency for Number-Correct Routing at Three Pass Rates.....	123
4.16 Decision Consistency for Random Routing at Three Pass Rates	124
4.17 Correlations Between True and Estimated Abilities at 30% Passing	125
4.18 Correlations Between True and Estimated Abilities at 40% Passing	126
4.19 Correlations Between True and Estimated Abilities at 50% Passing	127
4.20 Overall Root Mean Square Errors at 30% Passing	128

4.21 Overall Root Mean Square Errors at 40% Passing	129
4.22 Overall Root Mean Square Errors at 50% Passing	130
4.23 Routing Path Frequencies in 1-2-2 Design with Four Routing Strategies.....	131
4.24 Routing Path Frequencies in 1-3-3 Design with Four Routing Strategies.....	132
4.25 Routing Path Frequencies in 1-2-3 Design with Four Routing Strategies.....	133
4.26 Routing Path Frequencies in 1-3-2 Design with Four Routing Strategies.....	134

LIST OF FIGURES

Figure	Page
1.1 Three-Stage MST Design with 3 Levels of Difficulty in the 2 nd and 3 rd Stages	14
2.1 Illustration of Parallel Panel Structure	56
3.1 Test Structures of Interest	89
3.2 TIFs for Six Operational Forms and the Average TIF	90
3.3 Original Average TIF and TIFs Re-centered for Three Passing Rates	91
3.4 Target Test Information Functions for Three Pass Rates	92
3.5 Sample Assignment of Stage-Level Information Functions to Modules	93
4.1 RMSEs for DPI Routing with 1-2-2 Design at Three Pass Rates.....	135
4.2 RMSEs for DPI Routing with 1-3-3 Design at Three Pass Rates.....	136
4.3 RMSEs for DPI Routing with 1-2-3 Design at Three Pass Rates.....	137
4.4 RMSEs for DPI Routing with 1-3-2 Design at Three Pass Rates.....	138
4.5 RMSEs for Proximity Routing with 1-2-2 Design at Three Pass Rates	139
4.6 RMSEs for Proximity Routing with 1-3-3 Design at Three Pass Rates	140
4.7 RMSEs for Proximity Routing with 1-2-3 Design at Three Pass Rates	141
4.8 RMSEs for Proximity Routing with 1-3-2 Design at Three Pass Rates	142
4.9 RMSEs for NC Routing with 1-2-2 Design at Three Pass Rates.....	143
4.10 RMSEs for NC Routing with 1-3-3 Design at Three Pass Rates.....	144
4.11 RMSEs for NC Routing with 1-2-3 Design at Three Pass Rates.....	145
4.12 RMSEs for NC Routing with 1-3-2 Design at Three Pass Rates.....	146
4.13 RMSEs for Random Routing with 1-2-2 Design at Three Pass Rates.....	147
4.14 RMSEs for Random Routing with 1-3-3 Design at Three Pass Rates.....	148

4.15 RMSEs for Random Routing with 1-2-3 Design at Three Pass Rates.....	149
4.16 RMSEs for Random Routing with 1-3-2 Design at Three Pass Rates.....	150

CHAPTER 1

INTRODUCTION

1.1 Background

As computers have come to take on great prominence in many aspects of everyday life in recent years, so too has computerization come to the forefront of assessment practices at the outset of the twenty-first century. Tests from many testing programs now are administered exclusively by computer. For example, the Graduate Record Examination is a computerized test used in the context of admission to graduate school, the information technology field has the Microsoft and Novell certification examinations (among many others), and the Nurses Certification and Licensure Examination (NCLEX) is administered to thousands of prospective nurses annually. Many other testing programs are including studies of computer-based testing (CBT) in their ongoing research agendas (e.g., the American Institute of Certified Public Accountants, the College Board, and the National Assessment Governing Board).

While the trend toward computerization is certainly present in terms of educational testing, CBT is particularly becoming more prevalent in the area of professional certification and licensure assessment, as more and more credentialing agencies regard CBT as an effective mechanism for test delivery. There are a number of reasons for this, including that 1) an increasing number of professions are becoming more computerized, 2) examinees want to receive their scores more quickly, and 3) computerization of examinee responses facilitates data management. In addition, many professions are redefining and expanding the constructs they are trying to measure with such tests, and computers can give them added flexibility to obtain quality measurement.

With these sorts of general benefits associated with CBT relative to paper-and-pencil assessment for certification and licensure tests, the measurement advantages to be realized in operational testing do differ with respect to how a computer-based test is implemented (Drasgow & Olson-Buchanan, 1999). Some possible sources of variation include the choice of item type, scoring method, the relative inclusion of multimedia and other technological innovations in the test administration, the procedures for item and item bank development, and test designs. This last issue of test designs, sometimes discussed as test models, refers to structural variations in test administration. To be more specific, it addresses how the items in a test are sequenced and presented to examinees. Test design is a topic of much growing interest for research among test developers particularly given the evidence in the psychometric literature for improved measurement under adaptive test designs in CBT (Van der Linden & Glas, 2000; Mills et al., 2002).

The possibilities that 1) tests need not be exactly identical in sequence or test length and that 2) alternative designs could be implemented can be traced back to early work on intelligence testing done by Binet and Simon (1905, 1908). In these early tests, both starting and termination points varied across students and were dependent on the responses provided by individual examinees. From that work and later studies by many researchers including Lord (1970, 1971a, 1971b, 1971c, 1971d) came the notion of tailoring tests to individual examinees, and today the continuum of test designs used in practice with CBT ranges from linear fixed-form tests assembled well in advance of the test administration to tests that are adaptive by item or by sets of items and are targeted at the estimated ability of each examinee individually. Each of these designs possess a variety of benefits and drawbacks for different testing constructs, and making the choice

among such designs involves considerable thought and research on the part of a credentialing testing organization about the nature of the construct, the level of measurement precision necessary, and the examinee population.

Available test designs in the measurement literature fall into three categories, one that is not adaptive – linear fixed form test design, and two others that are adaptive – multi-stage test designs and computer-adaptive test designs. The first of these, the non-adaptive linear fixed-form test, has been widely implemented in both paper and pencil and CBT. In a CBT context, Parshall, Spray, Kalohn, and Davey (2002) described the linear fixed-form test as a computerized fixed test, or CFT. The second and third families of designs, multi-stage testing (MST) and computerized-adaptive testing (CAT), are both adaptive and are primarily implemented in a computer-based setting. There are substantial differences between these families relating to the test units by which the adaptive algorithm works: in the former case adapting to examinee ability occurs by sets of items while CAT is adaptive by individual items. These three families of designs are described in greater detail below.

CFT involves the case where a fixed set of items is selected to comprise a test form, and multiple parallel test forms may be created to maintain test security and to ensure ample usage of the item bank. In this approach, test forms may be constructed well in advance of actual test administration or assembled as the candidate is taking the test. This latter circumstance, commonly referred to as linear-on-the-fly testing, or LOFT, is a special case of CFT that uses item selection algorithms that do not base item selection on estimated examinee ability; rather, selection of items proceeds relative to other predefined content and other statistical targets (Carey, 1999). Each examinee

receives a unique test form under the LOFT design, but this provides benefits in terms of item security rather than psychometric efficiency, as noted by Folk and Smith (2002). Making parallel forms or introducing some randomization of items across forms can address item exposure and test security concerns. Some other advantages associated with linear fixed forms and LOFT include 1) the opportunity for examinees to review, revise, and omit items, and 2) the perception that such tests are familiar and easier to explain to candidates (Patelis, 2000).

The linear test designs possess many clear benefits for measurement, and depending on the purpose of testing and the degree of measurement precision needed they may be wholly appropriate for many certification and licensure organizations. However, other agencies may be more interested in other test designs that afford them different advantages, such as the use of shorter tests and the capacity to obtain more precise measurement all along the ability distribution and particularly near the cut-score where pass-fail decisions are made in order to classify examinees as masters or non-masters. The remaining two families of test designs are considered to be adaptive in nature, though they do differ somewhat with respect to structure and format.

The second family of test designs (MST) is often viewed as an intermediary step between a linear test and a CAT. As a middle ground, MST combines the adaptive features of CAT with the opportunity to pre-assemble portions of tests prior to administration as is done with linear testing (Hambleton, 2002a, 2002b). MST designs are generally defined by using multiple sets of items that vary on the basis of difficulty and routing examinees through a sequence of such sets based on performance on previous sets. As shown in Figure 1.1, for the more general MST design, each set of items an

examinee receives comprises a stage, and most of the most common MST designs use two or three stages, although the actual number of stages that could be implemented could be set higher (or lower) given the needs of different testing programs. In theory, with a sufficient item bank each of the sets of items administered in a given stage can be built to meet the specific statistical and content constraints of the test at large and yet vary by difficulty to ensure that the process of tailoring can proceed to a high level of measurement accuracy (and ultimately decision accuracy) for most candidates.

The third family of test designs, CAT, can be viewed as a special case of the MST model to the extent that CAT can be thought of as an MST made up of n stages with just one item per stage. In both cases the fundamental principle is to target test administration to the estimated ability of the individual. There are differences, of course: as item selection in CAT is directly dependent on the responses an examinee provides to each item singly, no partial assembly of test forms or stages takes place for a computerized-adaptive test prior to test administration. Furthermore, given that CAT is adaptive at the item level, Lord (1980) and Green (1983) indicate that this test design provides the most optimal estimation of candidate proficiency all along the ability continuum relative to other test designs.

However, there are limitations to the promise of CAT for credentialing assessment. One particular vulnerability of CAT from the perspective of examinees is the issue of item review (Wainer, 1993; Stone & Lunz, 1994; Vispoel, Rocklin, & Wang, 1994; Wise, 1996). Whereas in traditional paper-based administration examinees can go back and change answers as they see fit, this is not an option in most implementations of CAT because of the nature of the adaptive algorithm. Once an answer is provided to a

particular question and the examinee elects to go on to the next item, that response is used to determine the next item to be presented. If an examinee were allowed to return to previously administered items and change their responses to even a few items, it would limit the effectiveness of the process of adapting the test to evolving estimates of examinee proficiency. Given this difficulty, MST is an attractive alternative, because individual examinees can be given the opportunity to move around and answer items within a stage in whatever sequence they please. After completing a stage in MST, however, the items within that stage are usually scored using an appropriate IRT model and the next stage is selected adaptively, so no return to previous *stages* can be allowed (though, again, item review within a module at each stage is permissible).

Other potential difficulties associated with CAT from an implementation perspective include assuring proper content representation, the difficulty of using item sets where local dependencies may exist, the size of the item bank needed to support CAT while preserving low item exposure, and perceived inequities among examinees due to individuals receiving completely different sets of items. These drawbacks of CAT, when taken with the relative rigidity of linear test forms, promote continued investigation into alternative test designs within the broad heading of MST for certification and licensure assessment.

Thus, a primary distinction between test designs that can be made concerns the property of being adaptive or not. Traditionally, linear forms have predominated operational testing (both paper-and-pencil and computer-based). However, advances in research into item response theory over the years (Lord & Novick, 1968; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and the advent of

powerful and inexpensive desktop computers have facilitated implementation of adaptive test models. Such methods are described as adaptive in the sense that the sequence of items or sets of items administered to an individual examinee is dependent on the previous responses provided by the examinee (Lord, 1980).

To the extent that the item bank to be used is wide and deep enough to provide items that are maximally informative about each examinee's ability level (while maintaining item exposure levels sufficiently low enough to ensure test security), adaptive testing represents an approach to measurement that is more economical from an information-gathering perspective than the simple linear test form because the examinee's ability is factored into item selection. After administration of an item or a set of items on the test, the general methodology for adaptive tests is for the computer to use the item statistics computed under the principles of item response theory (IRT; see Hambleton, Swaminathan, and Rogers, 1991) and calculate a provisional ability estimate for the examinee; that provisional estimate is then used in to identify an item or set of items that the individual examinee will have on average a 50% chance of answering correctly. These are the test items for which predictions about each individual candidate's responses are most uncertain, and therefore, the most information about candidate ability is learned from administration of these items. Estimation of ability and administration of test items continues on in this way until some stopping rule is reached (such as presentation of a set number of items or the standard error of measurement for the examinee dropping below a pre-specified threshold).

Thus, adaptive testing represents a considerable step toward efficiency in measurement because an examinee that early on in a test exhibits high ability need not be

presented with many items of low difficulty, and conversely, a low-ability examinee would not receive many very hard items. With such efficiency, test length may also be reduced. Other advantages associated with adaptive testing include enhanced test security, testing on demand, individualized pacing of test administration, immediate scoring and reporting of results, and easier maintenance of the item bank (Hambleton, Swaminathan, & Rogers, 1991). At the same time, adaptive testing is administratively more complex, involves a changed approach to test development that is something of a departure from the procedures used in paper and pencil testing, and presents its own security concerns.

In sum, the choice of test design for a testing program is one that must be made with both measurement and practical considerations in mind. As both benefits and disadvantages of the different designs become clear through research, practitioners will be able to make appropriate decisions given the needs and peculiarities of individual testing programs. To that end, continued investigation into alternative test designs within the broad heading of MST for certification and licensure assessment is warranted.

1.2 Statement of the Problem

While three general families of tests designs exist, among paper-and-pencil (P&P) tests the linear test design is most commonly used, as in most cases implementing adaptive strategies for paper-based tests is not operationally feasible, although some recent research has explored this possibility (e.g., Zimowski, 1988,1989; Bock & Mislevy, 1988; Bock & Zimowski, 1989, 1998; Rock, Pollack, & Quinn, 1995). With

respect to CBT, operational testing programs have to this point for the most part implemented their assessments as either CFT or in a CAT format.

At once reflecting these trends and providing the empirical foundation for them, operational testing has by and large focused a great deal on linear fixed-forms and CAT with comparatively limited use of MST over the years, although the research base for the adaptive-by-stage testing method can be traced back over fifty years. Indeed, some of the initial research into tailored test methods was completed on tests that routed examinees through sets of items (rather than by each individual item) that varied by difficulty, and in recent years, MST has garnered increasing levels of interest by operational testing programs. This is made particularly evident by consideration of a number of important studies on MST that have been completed recently, including (but not limited to) Luecht, Nungester, & Hadadi (1996), Luecht and Nungester (1998), Patsula (1999), Patsula and Hambleton (1999), Reese and Schnipke (1999), Reese, Schnipke, and Luebke (1999), Schnipke and Reese (1999), Xing (2001), Xing and Hambleton (2001), Jodoin (2002), Jodoin, Zenisky, and Hambleton (2002), and Xing and Hambleton (2002) as well as the continued progress with testlet research (e.g., Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991). These studies represent important steps in exploring the psychometric properties of multi-stage tests relative to other test designs, but some research questions remain, particularly relative to MST-specific design variables such as how such tests are assembled in terms of structural variables and in light of specific examinee populations and item bank limitations.

In particular, the interaction between several specific design variables is not well understood in terms of measurement precision, decision accuracy, and decision

consistency as well as operational matters such as item and module exposure rates. Chief among these is the role of test information, including 1) the extent to which it is possible to decrease such information and yet still obtain high levels of measurement accuracy and 2) the distribution of such information across stages of the test. Two other issues of critical interest are test design structures and the routing strategy used. There are countless ways in which a multi-stage test can be structured with respect to both within- and across-stage dimensions, and the method used to identify which examinees are routed to which modules in stages after the first is a topic basic to the design. However, the MST literature to this point has been relatively ambivalent on advantages and disadvantages of routing strategies. Different authors have employed a variety of strategies to varying results, and no simulations or other operational studies have been undertaken to provide direct comparisons of different methods.

Before implementation of MST can continue on a large-scale, it is clear that work remains to be done to advance understanding of these issues and variables. To the extent that testing programs are looking to use MST for high-stakes decisions, investigation of the properties of different implementation strategies seems appropriate and useful to the measurement community.

1.3 Purpose of the Study

In focusing the research at hand to adaptive testing using such sets of items (or stages), this study is intended to build on the MST findings already in the literature to further understanding of the measurement properties of various MST strategies. This study involves an investigation of the relational impact of variables such as amount of

target test information, different passing rates, and routing strategies in the context of several commonly-researched MST designs. The purpose of this research is to advance an established line of research on the MST methodology by enhancing understanding of how such variables interact for estimating ability and ultimately impact pass-fail decisions for individual examinees.

1.4 Significance of the Problem

With respect to certification and licensure assessment, the purpose of testing is to identify those individuals who have met a particular set of standards within a specified profession, and as such tests used to grant professional competence need to be particularly precise in the area around the passing score (American Educational Research Association, American Psychological Association, and National Council for Measurement in Education, 1999). This condition provides a compelling psychological argument for some form of targeted or adaptive testing, to ensure that the maximum amount of measurement information is gathered from each candidate in order to make the most accurate decision reasonably possible in each case.

In a general sense, adaptive testing is defined by iteratively updating provisional estimates of examinee proficiency subsequent to receiving a response or set of responses from an examinee and then choosing the next item(s) based on the fresh estimate. Currently, adaptive testing methods are widely used for testing in a variety of educational and psychological contexts, although the most common implementation of the adaptive testing model in use today involves tests that are (1) constructed based on the principles of item response theory (IRT), (2) delivered to examinees by computer, and (3) adaptive

by item.¹ The NCLEX, used as a tool for granting professional licenses to nurses, is an example of one such test.

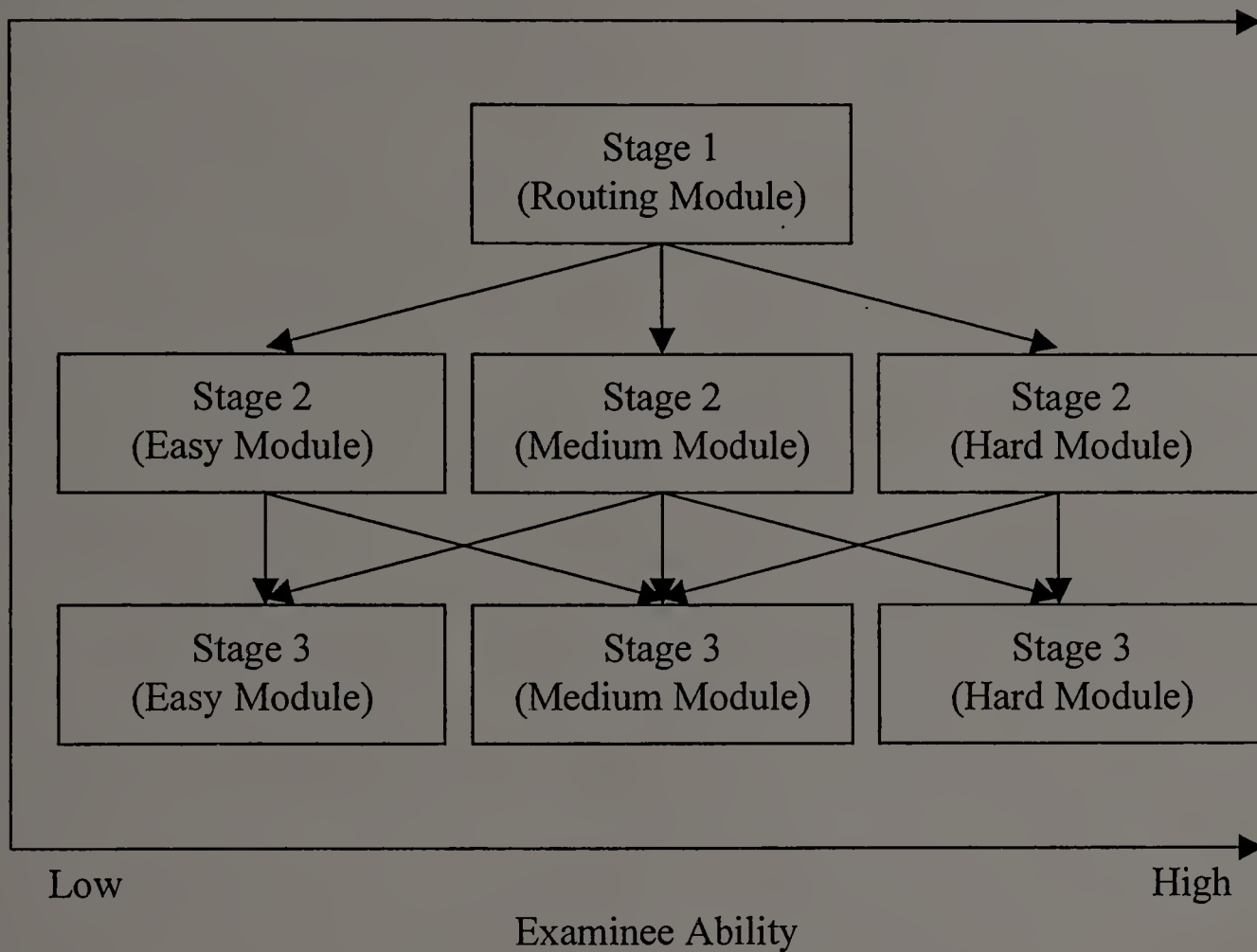
However, given that different test designs are differently appropriate in different testing situations, once the purpose for testing, the construct of interest, and the examinee population are taken into account, it is clear that research into alternative test designs such as variations on the MST model as described above is critical for informing the decisions of test developers who must be mindful of the needs of individual testing programs particularly in certification and licensure assessment. While the level of precision afforded a testing program by fully adaptive CAT may not be needed for examinees that clearly pass or clearly fail, professional testing programs may have other reasons for wanting to obtain more individualized measurement information for all examinees at all points on the ability scale. For example, a credentialing or licensure organization might perhaps be interested in providing diagnostic information for lower-performing candidates or publicly recognizing the performance of particularly high-ability candidates, in which case the additional measurement data obtained through adaptive methods would likely be regarded as a decided advantage for adaptive designs over traditional, linear, fixed-form approaches.

This study carries with it significant implications both for the theoretical underpinnings of the concept behind multi-stage testing and for operational psychometric practice. Ultimately, just as there are many ways of implementing CAT, there are many design variables that directly affect the efficacy and practicality of MST. As such, the

¹ That particular sort of assessment has been extensively researched and implemented (see edited books by van der Linden and Glas (2000) and Wainer, et al. (2000) for excellent overviews of the theory and practice in that area).

study described here represents an effort to clarify and further advance understanding of the psychometric properties of MST as it may be used in operational credentialing assessments.

Figure 1.1. Three-Stage MST Design with 3 Levels of Difficulty in the 2nd and 3rd Stages



CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Introduction

Provided in this chapter is an overview of the theory behind MST with particular focus on the basic structure of a multi-stage test. This includes information about variations on approaches to the MST model used in operational testing given various measurement and practical considerations, as there are a number of design variables that can have significant bearing on test results. Also, as important studies about the usefulness of adaptive-by-stage testing techniques have been completed in both IRT and non-IRT contexts, in summarizing the tradition of research into MST relevant research findings from both of those theoretical perspectives are detailed. Lastly, this review of the multi-stage testing literature concludes with the consideration of some different special-case applications of the multi-stage methodology and the highlighting of several areas for research that significantly inform the design of the current study.

2.2 Fundamentals of the MST Design

MST can be described as an approach to testing that involves the adaptive administration of sets of items to examinees. As such sets vary on the basis of difficulty, the particular sequence of item sets that any one examinee is presented with as the test is administered is chosen based on the examinee's ability estimate. After an examinee finishes each item set, that ability estimate is updated to reflect the new measurement information obtained about that examinee's ability through administration of the item set. In MST terminology, these sets of items have come to be described as modules (Luecht

& Nungester, 1998) or testlets (Wainer & Kiely, 1987)¹, and can be characterized as short versions of linear test forms where some specified number of individual items are administered together to meet particular test specifications and provide a certain proportion of the total test information. The individual items in a module may be all related to one or more common stems (such as passages or graphics) or be more generally discrete from one another, per the content specifications of the testing program for the test in question. These self-contained, carefully constructed, fixed sets of items are the same for every examinee to whom each set is administered, but any two examinees may or may not be presented with the same sequence of modules.

The stage in multi-stage testing is an administrative division of the test that facilitates the adapting of the test to the examinee, and each examinee receives a minimum of two stages' worth of modules (the exact number of stages is a decision for test development relating to content coverage and measurement precision). In each stage of the test, an examinee receives a module that is selected as appropriate for that examinee in terms of difficulty based on the ability estimate computed from performance on the stage(s) prior. Within a stage, there are typically two or more modules that vary from one another on the basis of average difficulty. As candidates progress through the test, they are routed to the one module within each stage that is likely to be most informative for estimating that individual's true ability: strong candidates receive modules of higher average difficulty, while less able examinees are presented with modules that are comparatively easier.

¹ While testlet is sometimes taken to refer to a set of items linked by a common stem or otherwise dependent on one another, it has more recently referred more generally to any set of items designed to be administered as a group within a larger test instrument. As such, though module is used within this research for consistency, the terms module and testlet are interchangeable in the literature.

A typical administration of a multi-stage test, constructed and administered in an IRT context, proceeds as follows. An examinee in the first stage is typically administered a module of medium difficulty. As with many applications of adaptive testing, the use of a medium difficulty starting point is common because no prior information about individual candidates is known. As a starting point in MST, medium difficulty modules are likely to be informative from a measurement perspective for a large proportion of candidates and allow for highly efficient routing of many candidates to second-stage modules. After the first-stage module, the examinee's responses are scored and a provisional ability is estimated using one of a variety of methods, primarily Bayesian or maximum likelihood estimation (Hambleton, Zaal, & Pieters, 1991). The examinee is then routed to a second-stage module based on that estimated ability, and this process continues through as many stages/modules as the testing program deems necessary to achieve a desired level of measurement precision, decision accuracy, or test length.

In terms of implementing MST, as with CAT, there are many design variables and development procedures that come together and impact what the finished product of a multi-stage test's 'test form' looks like both psychometrically and from the perspective of the examinees. One of the advantages of the multi-stage design is that there are numerous ways in which it can be implemented and so it is a highly customizable design for testing programs to use. At the same time, many of the practical issues that arise with CAT (as inventoried by Green, Bock, Humphries, Linn, and Reckase (1984), Mills and Stocking (1996), and Wise and Kingsbury (2000)) are directly relevant to implementation of MST. However, the issues that occur with the development and operational use of MST are different enough to warrant a review of the design variables present in MST, the

methods used in developing multi-stage tests, and the operational issues that must be reckoned with, given that the properties and features of MST for certification are continuing to evolve and emerge as research goes on.

For example, in providing an overview of two-stage testing using IRT, Lord (1980) outlined a number of design considerations that he identified as impacting the nature and quality of ability estimation from tests using a two-stage procedure. His ideas, as abstracted below, can be generalized to a test of n stages:

- Total number of items in test
- Number of items in initial and each n -stage module
- Difficulty of the initial module
- Number (and difficulty) of alternative modules in each n -stage
- Cut-points for routing examinees to modules, and
- Method for scoring stages and each n -stage test.

While Lord suggested that it was not possible to identify truly statistical optimal designs for each and every operational testing context, it seems entirely reasonable to find combinations of these variables that would provide high-quality results as needed for a particular test's use or the interpretations to be made based on the test scores.

To Lord's (1980) list can be added several additional considerations that have emerged through MST research, including the number of stages, the ability distribution of the candidate population, the extent of target information overlap for modules within stages, whether random module selection (at appropriate difficulty level) or panel-based administration is used, whether content-balancing is done at the module or total test level, choice of method for automated test assembly, the size and quality of the item bank, how

test information is distributed across stages, placement of cut-scores for pass-fail decisions, the issue of item review, and item exposure levels. To facilitate understanding of the issues involved, each of these considerations can be loosely clustered as related to either (1) basic structure variables, (2) test and module assembly issues, or (3) administration. Each of these clusters and its associated variables are defined and detailed below.

2.2.1 Basic Structural Variables of the MST Design

There are several MST design variables that, when taken together, help to define the basic structure of an MST in practice. The first of these is the total number of items in the test. An often-cited benefit of adaptive testing is the opportunity to shorten tests in terms of the number of items presented to each examinee (thereby reducing testing time) by targeting tests to examinee ability, a test that is adaptive either by items or by stages need not necessarily be as long as a linear test form (Bergstrom & Lunz (1992), but considerations of domain coverage and measurement precision still must be balanced.

Research in MST specifically in a certification context has seen a wide range of test lengths, including studies with over 150 items administered to examinees over six stages (Luecht & Nungester, 1998) and with 35 items (two stages), as found in some information-technology testing applications (Xing & Hambleton, 2002). A recent study by Jodoin, Zenisky, and Hambleton (2002) found that a 40-item two-stage test performed nearly as well as a 60-item three-stage test (as represented by decision accuracy (DA), kappa, and correlations between true and estimated abilities from each design). In both cases the number of items per stage was held at 20. The key point to be made here is that

as compared to a fixed-length exam, multi-stage tests can be shorter, although the exact reduction in the number of items is a matter of both research and practical considerations such as content coverage.

When describing test length in the context of MST, while that quantity is clearly defined by the total number of items in the test, it also concerns the total number of stages in the test. The MST literature itself is divided on how many stages provide optimal measurement, and as with many of the other design variables in MST that number is closely related to other issues such as how many items are to be included per stage. While most of the MST research to date has focused on two- and three-stage tests in which all examinees receive the same number of stages, there are exceptions, of course. The literature on computerized mastery testing (CMT) which is a variation on the basic MST approach involves variable-length mastery tests where different examinees may receive different numbers of modules, and a four-stage test was the focus of a study by Luecht, Nungester, and Hadadi (1996; also Luecht & Nungester, 1998). The number of stages is also affected by policy considerations: for example, in a high-stakes context, stakeholders may not be comfortable using a two-stage test due to a perception of some candidates being unable to recover or 'pass' if their true abilities are at or above passing and they are routed to a lower-difficulty module in the second stage. Clearly, measurement efficiency is not the only consideration taken under advisement in the process of deciding the appropriate number of stages to include.

Along with establishing the total test length and the number of stages, another critical consideration is of how many items per stage to administer. If ease of explanation to candidates and greater standardization of module development is a priority for a testing

agency, then it may be preferable for modules within and across stages to be of equal length. Some recent studies (Jodoin, 2002; Jodoin, Zenisky, & Hambleton, 2002; Hambleton & Xing, 2002) have implemented modules consisting of 20 items in each of three stages, while Luecht and Nungester (1998) worked with three-stage tests composed of modules that were 60 items in length (total test: 180 items). Alternatively, work by other researchers has explored other configurations of items, such as longer first-stage tests (Xing & Hambleton, 2002) or tests with more items in the stage(s) after the first (Loyd, 1984; Reese, Schnipke, & Luebke, 1999; Schnipke & Reese, 1999; Reese & Schnipke, 1999; Kim & Plake, 1993; and Castle, 1997). Patsula (1999) defined the rationale for longer first stages as relating to the need for more accurate measurement in the first stage prior to routing (the 'Routing Test' strategy), while extending the length of subsequent stages may be justified by the thinking that since the tests are more closely aligned with examinee ability at later points in the test, providing more items tailored to estimated ability in those stages is capitalizing on the information obtained from candidates after some routing has been done (the 'Higher Stage' strategy).

One caution to module length mentioned in the literature is the need to keep module length consistent within stages. This is to say that for reasons of fairness testing programs may want to avoid routing some examinees between stages on the basis of more or fewer items than other examinees. This caution does not preclude longer first stage tests followed by shorter modules in subsequent stages (or vice versa) but is just to endorse uniformity for fairness within stages (Luecht & Nungester, 1998).

2.2.2 Test and Module Assembly

The characteristics and methods involved with the actual construction of MST tests and modules are among the design considerations in MST that are particularly complex and which most clearly help to differentiate MST from other test designs. For example, the difficulty of the first-stage module in an MST is a critical decision of test development. The choice of a starting point for a multi-stage test is much the same as it is for CAT: namely, in the absence of information about an examinee, the optimal starting point is in the area of medium difficulty to obtain maximal information about as many candidates as possible at the outset. Thissen and Mislevy (2000) suggest for test developers to stipulate an initial estimate of ability that specifies what difficulty level of testlet to begin with. The notion of maximizing information from the very start in CAT is an even greater necessity in MST because the adaptive routing does not occur until an examinee has already received perhaps as much as one-third or even half of the test via the first stage. For this reason, in the context of testing with a relatively normal distribution of candidates, starting with a medium-difficulty module helps to ensure that the initial module presented will be informative in a measurement sense for a large proportion of examinees.

Once examinees have been administered the first-stage, the critical issue of the MST design concerns the number of and the relative difficulty of the modules in each and every subsequent stage, an issue discussed at length by Lord (1980). In the prototypical MST presented earlier in Figure 1.1, there are three modules in both the second and third stages. The design process for these modules in stages subsequent to the first is contingent on several points, including the level of routing precision desired by the

testing program, the depth and breadth of the item bank, and the extent to which such modules should be discrete or can overlap. Notice that these differ by difficulty: for example termed easy, medium, and hard, they are generally aligned relative to the ability continuum of candidates, such that lower-ability candidates should be routed to the easier modules in each stage while more proficient examinees would be presented with more difficult modules.

However, to the extent that easy, medium, and hard are relative terms that have meaning for items, these modules are actually referenced by the ability scale (which in IRT generally ranges from -3.0 to 3.0). An example in the case of a stage with three levels might involve using test assembly procedures to target the three modules at -0.5 , 0.0 , and 0.5 , respectively. In the process of constructing such modules, a testing program might want to make the modules more distinct from one another, such as in recent studies by Jodoin (2002) and Xing and Hambleton (2002) where the easy and hard modules were transformed by one full standard deviation.

However, as module difficulty is generally defined by average b -parameter estimates, such averages can be obtained in two ways. Lord (1980) referred to these as either peaked or non-peaked distributions of items within modules. Peaked modules are those in which items are all of approximately equal difficulty, while non-peaked modules contain more variation and so the average difficulty is arrived via a more heterogeneous assemblage of items.

Another consideration in the development of an MST is whether content-balancing should occur at the module or total test levels. As described previously, the absolute number of items is dependent on the complexity of the construct of interest.

However, a related point for test developers exploring MST as a test design is whether domain coverage should be achieved within stages or across the whole test (Luecht & Nungester, 1998; Folk & Smith, 2002). To meet elaborate content specifications within stages can require more items at each stage, while meeting test specifications across an entire test provides greater flexibility in terms of test assembly.

A difficulty in content-balancing across the entire test, however, is that test users may not consider routing examinees through a limited number of stages when each of the stages is not reasonably representative of the domain of interest appropriate. In other words, if the set of items an examinee is given only covers a portion of the test specifications, can decisions about the rest of the test be based on data that is incomplete in that respect from a fairness perspective? Research is not clear on this point, but it may be that stages with fewer items in relatively constrained domains of interest (i.e., reading comprehension) may be perfectly appropriate for content-balancing within stages whereas more content-based and/or cognitively complex domains may require more items within a stage to accomplish the same goal. In some testing applications, resolving this dilemma may result in the administration of more items than are strictly necessary for precise ability estimation (Folk & Smith, 2002).

Research into item bank size and the quality has been extensively studied in the specialized context of CAT, but empirically speaking this topic is only now beginning to be considered specifically in the context of MST, particularly as advances are made in the area of automated test assembly. A notable exception concerns recent studies by Xing (2001) and Xing and Hambleton (2002). In the Xing study (2001), varying conditions of item bank size and quality and placement of passing score were compared for a CAT, a

two-stage test, a three-stage test, and a linear form. Of the 72 possible conditions in the study (4 CBT designs x 2 levels of bank size x 3 levels of item quality x 3 levels of passing score), it was found that as item quality improved so did both decision accuracy (DA) and decision consistency (DC). In addition, larger item banks were found to be preferable. Also, because of the potential for lowering exposure levels of item and increasing test information, Xing noted that the benefit of larger item banks came in the form of greater ability to meet statistical targets such as test information functions and automated test assembly constraints. A subsequent study further exploring variations in item bank size and item quality (Xing & Hambleton, 2002) found little difference among different test designs (linear forms, 2-stage MST, and CAT) but the quality and size of the item bank did make a practically significant difference in the results.

To implement many of the design variations, item bank considerations are critical in that automated test assembly builds require that the item pool be of a depth and breadth to support such construction. For certification and licensure programs looking to move from paper-based to computerized-adaptive formats (such as CAT or MST), the item bank may have to be augmented in a substantial way because in non-adaptive pass-fail testing, items are often targeted to the cut-score. Given the focus of test development in credentialing on accuracy of pass-fail decisions rather than maximal information at the ends of the ability scale, there will likely be a relative dearth of items at the easy and hard ends of the difficulty scale. Therefore, test development for MST may be hindered in the process of assembling varying difficulty modules.

The ability distribution of the candidate population is a matter of importance in test development with any test design, but is a particular concern with the development of

MST. As explained in a study by Hambleton and Xing (2002) that included this as a variable, where the candidate population is located in the ability continuum has clear implications for measurement with regard to where target information functions are centered. This issue is one of efficiency: of course, cut-scores may be set independent of the nature of the candidate ability distribution, but the characteristics of the candidate pool have an effect on the process of module and test assembly (such as regard for average module difficulty and discrimination). Also, the shape of the ability distribution can impact module exposure, depending on how and where the cut-scores for routing are placed.

Similarly, deciding how to distribute test information across stages involves weighing efficiency and using test design to maximize the information to be obtained. This notion of using test information in the development of tests in a panel-based structure has been described by Luecht (2000) as a way to provide consistent control over error variance of estimated scores at various regions of the proficiency scale, in contrast to CAT where the 'target' test information function (TIF) can be understood as the overall maximum information possible after the last item is administered to an individual examinee (for maximizing score precision). For MST, however, modules can be viewed as intermediate administration structures of the test, and thus TIFs are specified for each module. The issue in this attribute of the MST design focuses on the partitioning of the target test information function across stages: is it better for measurement to obtain greater test information early on in the test, or hold off and wait until some tailoring of the test form has taken place? This is an important area for research.

In large part, the literature on methods for automated assembly of modules and tests for MST builds on the extensive psychometric research that exists for item selection and test assembly for CAT, but automated test assembly (ATA) in an MST context is an aspect of the design that contributes substantially to differentiating MST from the other test designs. ATA software is designed to implement optimization algorithms or heuristics (or both) to satisfy certain content or statistical goals and explicit and implicit rules about test fairness and test content (Wightman, 1998), and it is all done in advance of testing, which permits human review of the modules. This systematization allows for the process of module development to be more standardized, particularly with respect to difficulty and test information, and reduces the labor-intensive task of hand-assembling the numerous modules needed for a large-scale, operational, high-stakes MST testing program. ATA software in effect requires that the constraints and goals of the modules to be built be specified as a mathematical optimization model made up of an optimization model to be maximized or minimized (Luecht & Nungester, 1998), and the task for the software is to solve that model using linear programming, network-flow, or some other such approach.

In the context of assembling a multi-stage test, the issues are many: as described by Luecht and Nungester (1998), the challenges include the potential to have the algorithm simultaneously solve more than one objective function, the possibility of different specifications for different modules, and the need for multiple replications to ensure module security and minimize item exposure. Another consideration for ATA is related to test structure: multi-stage tests can be built and balanced at either the level of modules (bottom-up) or in the total test administered to each candidate (top-down). This

distinction is critical in terms of identifying how many target TIFs are necessary for the automated test assembly 'builds' and thus, with MST test assembly being described by Luecht and Nungester (1998) as either a bottom-up or top-down enterprise, the terminology invokes a useful set of visuals for conceptualizing the ATA process.

A significant logistical issue with the development and administration of a multi-stage test concerns whether stratified random selection of modules (at appropriate difficulty level) or panel-based administration is used. An issue peripherally related to the decision of content-balancing within or across modules is whether to establish a bank of modules at the requisite different difficulty levels to be selected arbitrarily as candidates move from stage to stage or to use what is referred to as multiple panels (Luecht and Nungester, 1998; Luecht, 2000). A panel can be conceptualized as a specific and fixed set of modules that is assembled before administration and consistently administered as a fixed group (Figure 2.1). In a panel, the possible pathways that any one examinee may be routed through during the course of administration of a multi-stage test are identical for every examinee receiving that panel. Just as multiple parallel forms are made for linear tests, for reasons of test security, multiple panels that are developed to be classically parallel may be constructed and used in MST. To be clear, in the panel structure each panel represents the complete set of unique paths through n stages' worth of modules that an examinee may take, but examinees are not commonly routed through multiple panels.

In contrast, stratified random module selection involves modules that are constructed to be more discrete units, and these modules can be assembled in any order and combination during the course of the test. Whereas the panel structure is defined by

parallelism of the panels, this method is predicated on parallel modules because of the process of random module selection. This approach does not lend itself to advance checking of pathways.

With respect to test assembly, there are advantages to each approach. The panel structure gives test developers control at the 'front end' in terms of managing pathways and ensuring that the complete tests are representative of the test specifications, while random selection requires that such controls be built into the 'back end' of test development as they are utilized during test administration. In cases of potential compromise, however, testing programs might consider it preferable to be able to pull out individual modules rather than remove entire panels from active administration. With respect to test security, a testing program may have many modules created to be parallel at each difficulty level, or alternatively tens or hundreds of parallel panels. In either case, test developers can activate as many or as few modules and/or panels as are needed.

With respect to the process of creating tests using any test design, one additional important consideration for test developers concerns item exposure and test security. In CFT, the parallel forms are used to help ensure a measure of security in terms of controlling item exposure by limiting the number of examinees who are presented with any one form, while in CAT complex item selection algorithms are used to promote usage of the entire bank. With respect to MST, creating multiple parallel modules at each difficulty level and developing numerous parallel panels helps to address these issues. In addition, different decision rules for routing can distribute examinees differently to modules in ways that lessen or increase item exposure.

2.2.3 Administration

With MST modules and (if used) panels in hand, there are a number of decisions relative to MST administration that must be made that impact the efficacy and implementation of MST. One consideration for implementation of MST (and other test designs) is whether or not to permit examinees to return to previously administered items during the course of testing. In CAT, of course, this is not a practical option due to adaptive routing decisions (although recent work by Papanastasiou (2002) explored a rearrangement procedure for adaptive testing with review), and in linear testing item review does not negatively impact any routing of examinees. The various studies that have been completed to determine any empirical beneficial or detrimental effects of review in adaptive testing have returned mixed results. For example, while a study by Lunz & Bergstrom (1994) found no significant ability differences among examinees that did and did not use review, they did find that simply being allowed the review opportunity resulted in significantly better scores.

In most MST applications, the decision to permit examinees to complete items in most any sequence of items within a module is trivial: however, the decision to permit review between modules encounters the same obstacle as is found between items in CAT. For this reason, review within stages is generally permitted, but not across stages. Ultimately, research seems to suggest that the primary benefit of item review is related to a psychological comfort factor, and in the context of certification and licensure using MST, the option to review within stages may serve to alleviate anxiety for some candidates (Patsula, 1999).

Lord (1980) cited the issue of strategies and cut-scores for routing examinees to modules as particularly critical, as the quality of the method by which examinees get routed to certain modules as opposed to others defines the usefulness of an adaptive, multi-stage administration. Some of the options cited in the literature for routing examinees to modules between stages include using number-correct (NC) scoring, cumulative weighted number correct, and IRT-based provisional proficiency scores such as maximum likelihood estimates (MLE) or estimated *a priori* (EAP) estimates (Luecht, 2000; Armstrong, et al., 2000; Wise, 1999). Other approaches also considered in the literature include using maximum testlet information and Wald's (1947) SPRT (Luecht, Nungester, & Hadadi, 1996). To implement number-correct scoring, Luecht, Brumfield, and Breithaupt (2002) suggested incremental computation of upper and lower bounds for NC scoring of various combinations of routings through the panel structure. Location of routing points can be done using either the approximate maximum information (AMI) or defined population intervals (DPI) approach. The AMI method uses cumulative TIFs to identify optimal decision points for module selection, while the DPI structure is used to specify proportional routings through the panel and module structure.

Through Lord's (1980) research, he suggested that the difficulty levels of the modules should match the estimated ability levels of the candidates who are routed to them. As his work was based on trial-and-error efforts in setting cut-scores, he recommended that the topic of empirically derived cut-scores deserved more study, and this today remains an important area for investigation. In terms of other recent research into this topic, Schnipke and Reese (1999) used number correct scoring and a simulation study methodology to try and determine cut-scores for classification at each level. Their

approach involved trying to minimize mean-squared error (MSE) of theta estimates from simulated examinees administered easy, medium, and hard modules in turn to figure out at which number-correct value MSE was lowest between low and medium modules and medium and hard modules. A study by Thissen (1998) explored a variation on a fixed-weight scoring method for testlets that allowed for standard errors of ability estimates to be available. Dodd and Fitzpatrick (2002), in discussing the Schnipke and Reese (1999) and Thissen (1998) studies, related the two in the context of advancing a routing method that is both number-correct and information based. This recommendation involved computing number-correct theta estimates and then selecting modules based on information at that estimate.

Kim and Plake (1993) used a simple comparison procedure in which examinees were routed to the module whose average difficulty most closely matched their estimated ability on the ability scale. Hambleton and Xing (2002) chose to implement strategies anchored to the proficiency scale (related to the DPI method suggested by Luecht, Brumfield, and Breithaupt, 2002). Here, approximately equal numbers of candidates were routed to each second level module. A suggested variation on this approach is to have examinees within two standard errors of the value that the MST is targeted at routed to the middle difficulty module; examinees on either side of those cut-of values are routed to the easy or hard modules as appropriate.

Another aspect of routing concerns the possible pathways for routing (Luecht & Nungester, 1998). To the extent that examinees are routed between modules from stage to stage, the number of possible pathways for routing is a variable that can also be controlled by the testing program. In some testing applications, examinees might not be

permitted to move from the easiest module in one stage to the hardest module in the immediately subsequent stage. Such dramatic changes in estimation of ability between later stages are not likely under normal testing conditions (Luecht & Nungester, 1998), and may well be considered a flag for score review for some testing programs.

Closely connected to the methods for routing are the methods for scoring modules and the entire test. Lord (1980) suggested that in a situation with statistically equivalent items, simple number-correct scoring could be appropriate. In the psychometric literature, while relatively few studies have focused directly on this aspect of the design, scoring in the context of MST has involved Bayesian analysis, approaches based on maximum-likelihood estimation, the testlet models of Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (2000), and more extensive methods based on number-correct scoring.

Schnipke & Reese (1999) authored an important study that explored the use of number-correct routing and Bayes modal estimates of ability in the context of two-stage, multi-stage, and maximum-information testlet-based designs. Thissen (1998) obtained EAP ability estimates for candidates based on a pattern of two or more summed scores, and also developed a method for using Gaussian approximation to EAP ability estimation that is in essence a weighted linear combination of such estimates from separate summed scores, which allows for the estimation of ability from raw score patterns obtained through MST.

2.2.4 Summary of MST Design Variables

Clearly, there are many design considerations for the development of MST that can significantly affect what a multi-stage test looks like and how the results obtained fit in with the purposes and goals of a particular testing program. A multi-stage test as used in practice can run the gamut of possibilities from resembling a linear test by implementing just a few very long modules and stages, to resembling a CAT with many short stages. Fortunately, as described in the next sections, the research base for information about MST is considerable, and developing understandings of the 1) psychometric issues that arise in the construction and administration of MST and 2) relationships between such design variables are both particularly active areas for MST research of late.

2.3 Foundation Research in Adaptive-By-Stage Testing

As research into assessments that were adaptive by item or by stage got underway in the early-to-mid twentieth century, these kinds of exams became known as programmed or branching tests. This was owing to the evolving nature of the relationship between examinee responses and the selection and presentation of subsequent items or stages. Other authors termed such tests tailored, in that the item selection was fit to current best estimates of examinee ability (Turnbull, 1951, as noted in Lord, 1980).

In Binet and Simon's (1905, 1908) studies involving paper and pencil intelligence testing, they developed a series of thirty individually administered assessments in which tests and items were arranged sequentially (in ascending order of difficulty) with the

understanding that students who were unable to answer easy questions would, in most circumstances, be unlikely to correctly answer more difficult items. Later decision theory work of Wald (1947) also factored significantly in advancing this concept of incorporating data into the estimation process as it became available, the objective that underlies adaptive data analysis methods. Wald's efforts described an approach to classification employing a sequential probability ratio test (SPRT) that involved hypothesis testing during data collection, the end result of which in testing environments is a classification decision (e.g., pass/fail, master/non-master, certified/not certified). Generally speaking, in sequential analysis, as data is collected on a case (or, in an assessment context, an examinee), information is compared with certain threshold values and conclusions about the case are iteratively updated until some stopping rule is reached.

The theory behind these first techniques for adaptive testing clearly had application in the context of educational and psychological testing (Krathwohl & Huyset, 1956; Patterson, 1962; Ferguson, 1969a, 1969b), and was also advanced in the area of personnel decisions through research by Cronbach and Gleser (1965). The U.S. Army's Behavioral Science Research Laboratory further extended research into branched tests with the work of Bayroff and Seeley (1967), and Bayroff, Ross, and Fischel (1974) by examining the comparability of adaptive and linear test forms. Lord's (1971a, 1980) description of the flexilevel test also represents a variation on the branching design. The flexilevel test is non-computerized approach in which test forms are printed with items ordered by difficulty: each answer the examinee provides is either right or wrong and the examinee is directed to follow the rule of responding to the next harder or easier item on that basis.

Early studies of tests that were adaptive by stages incorporated mechanical branching rules independent of IRT with paper and pencil tests. Angoff and Huddleston (1958) explored the possibility of a two-level testing system for the College Board's Scholastic Aptitude Test (as it was then known), finding that administrative complexities curbed the benefits of increases in reliability and validity that could be gained through use of two-stage testing, while Cronbach and Gleser (1965, chapter 6) studied the idea of two-stage testing in personnel decision-making.

Using classical test theory, Cleary, Linn, and Rock (1968a) developed four methods of constructing programmed tests in two-stage testing, which varied from one another in terms of how examinees were routed from the first to the second stage of testing. They termed the initial set of items provided to examinees the routing test, which was followed up by a measurement test. Their work involved simulation of the different routing conditions using 11th-grade student item response data for School and College Ability Tests (SCAT) and Sequential Tests of Educational Progress (STEP), and in this study they found that 40-item tests from each of the four of the routing methods looked at provided results (in terms of reliability) that were very much comparable with the full 190-item linear test, although using only the 40 most discriminating items for a CFT from the 190-item bank provided equally reliable results. In a subsequent study, Cleary, Linn, & Rock (1968b) used the same data to expand on one of the four methods in the earlier study, finding that on average a 40-item two-stage programmed test applying what they called a three-group sequential method again provided results that were quite comparable with the 190-item CFT.

A third study by these researchers (Linn, Rock, & Cleary, 1969) continued the previous studies by focusing on the five strategies for routing examinees previously considered as well as two additional methods. In this case the evaluative criteria was not only the internal criterion of total test score on the 190-item test but also the external criteria of scores on two Preliminary Scholastic Aptitude Tests (Verbal and Mathematics) and two College Board Achievement Tests (American History and English Composition). Against the internal criterion, all of the programmed methods performed well, although the authors noted that results from a shortened 40-item linear test with highly discriminating items were sufficiently comparable. However, with the external criteria, four of the programmed testing methods evaluated (and the group-discrimination method in particular) did actually exhibit higher correlations than the linear tests. It was further recognized that the relative simplicity of the group discrimination method meant that it could be most readily implemented in paper and pencil testing relative to all of the other methods studied.

As this research into two- and multi-stage testing based on classical test theory was underway, Lord and Novick (1968) were outlining the fundamental tenets of modern test theory. This represented a tremendous step forward for testing and adaptive test methods, as with the advent of IRT it was possible to simultaneously incorporate more item information beyond just the difficulty parameter, as was done in most of the earlier studies into branched tests (Kim & Plake, 1993). Furthermore, comparisons between examinees taking different items could be more easily made under IRT given the property of invariance for item and ability parameters. Using IRT as the basis for adaptive testing also provides improved measurement precision, the potential to

maximize testing efficiency for each examinee (given a sufficiently broad and deep item bank, to ensure that as many items targeted at the examinee's ability level as possible can be administered), shorter tests, suppression of omitted response options, and enhanced test security (Lord, 1977; Hambleton, Swaminathan, & Rogers, 1991). Clearly, there are many advantages to adaptive testing strategies that are well documented and recognized in the psychometric literature, and such benefits have become particularly evident in light of widespread understanding and implementation of IRT.

One of the first authors to provide a framework and measurement justification for adaptive-by-stage testing with IRT was Lord (1971b), who described two-stage testing as a method for providing the advantage of improved measurement for not only typical examinees but also those at the extremes of the ability distribution. In one design proposed, items were assumed to differ only with respect to difficulty, though within each individual second-stage test items were of more or less equal difficulty. In the other design, the difficulty of the second-stage tests are overlapping and each test within a stage (referred to by Lord as levels) should be maximally efficient for assessing examinees in its part of the score scale, while being economical about usage of the item bank. Simulation results suggested that just three or four levels of the second-stage test provided reasonably good measurement results. These results were consistent with later findings from studies of the self-routing multilevel test (Lord, 1971b, 1974, 1980; Marco, 1977), although information at the ends of the ability distribution was slightly lower than as was seen in the middle range.

An additional program of studies into adaptive-by-stage testing was undertaken by researchers at the University of Minnesota in conjunction with the Office of Naval

Research, in which the measurement properties of two-stage testing were evaluated for possible use in military testing. Betz and Weiss (1973) found that relative to conventional linear tests, the two-stage design they were studying had lower rates of misclassification (4%-5%) and that scores from the two-stage test were somewhat more variable on average. In a follow-up study designed to be a generalization of the 1973 study, Betz and Weiss (1974) implemented two two-stage strategies where the first was as before but the second consisted of a routing test that was somewhat harder and a second-stage test with items that were more discriminating on average. The findings from this study showed that recovery of the true ability distribution was best with the improved two-stage design, and reliability of the second two-stage test was higher. These results are significant in that they highlight the relationship between the item statistics of the items grouped together in the modules and the quality of the measurement obtained. Similarly, a subsequent study by Larkin and Weiss (1975) reported that as examinees were more accurately routed to second-stage tests by improved first-stage tests, misclassification decreased substantially.

2.4 Current MST Research

Clearly, given the design variables detailed previously, what is generically referred to as 'the MST design' in fact comprises an enormous range of theoretical and practical alternatives for implementation. While these variations do correspond to a high level of complexity for implementation, this design also represents tremendous flexibility for individual agencies. With such an accommodating design, MST is a very customizable approach to obtaining measurement precision for examinees along the

ability continuum. However, the measurement properties associated with the many possible MST variations are not yet well understood, and so research into applications of MST using IRT has continued in three primary directions.

First, many studies have taken an outcomes-oriented approach with particular focus on the effects of various test structures and different implementation strategies, particularly with respect to the dimensions suggested by Lord (1977, 1980). Comparing results from simulation studies of MST and other test designs has been a primary goal of such research, with two outcomes of special interest to researchers with regard to measurement with MST: the first of these concerns the quality of ability estimation across the continuum of ability and the second involves evaluating the efficacy of the design for classification of individuals into pass-fail categories.

A second critical area for current MST research is the development of ATA algorithms for assembling forms. As computer processors become more flexible and better able to simultaneously consider multiple constraints for building modules, a number of researchers are capitalizing on such power to produce increasingly complex methods to meet the many such constraints quickly and efficiently.

The third direction for MST research to this point has been on investigating several specialized cases of the MST methodology within the family of MST designs. Applications of MST that involve non-adaptive stage selection techniques and deal with the situation of local dependence within modules represent important variations on the basic MST model that are deserving of empirical investigation. In this section, the state of current research relating to each of these areas will be detailed.

2.4.1 Evaluating MST Relative to Other Test Designs

Numerous recent studies of MST involve examination of the quality of ability estimates with respect to the entire continuum of candidate ability, where criteria such as root mean square error (RMSE), bias, and relative efficiency are used to compare true and estimated values for simulated candidate ability. To evaluate ability estimation across the entire ability scale, RMSE provides a measure of accuracy between true and estimated ability values by computing the square root of the mean squared difference between those values at different ability levels on the θ scale. Bias refers to the difference between the mean of the estimates and the true ability at various levels of θ (when bias is positive, ability has been underestimated; conversely, negative bias is indicative of ability overestimation). The last index commonly used in such studies is relative efficiency, which provides a mechanism for comparing average standard errors from different test designs at different ability levels.

In the work of Reese and Schnipke (1999), where the efficiency of a two-stage testlet design was compared with CAT and a paper-and-pencil linear test, ability estimation was evaluated using RMSE and bias. Across the entire ability distribution, the CAT naturally exhibited the lowest RMSE and the least bias, although the most carefully constructed two-stage tests were actually the most error-free in the ability range from -2.0 to 2.0 . A subsequent study by Reese, Schnipke, and Luebke (1999) that focused on strategies for optimal assembly of testlets found that a carefully constructed and content-balanced two-stage test outperformed the CAT and the paper-and-pencil test in the middle portion of the ability scale with respect to both bias and RMSE, even though the statistical constraints for assembly were not strictly met. An additional study authored by

Schnipke and Reese (1999) found that several testlet-based designs (including a basic two-stage design, a two-stage design with the possibility of changing second-stage levels if misrouting was suspected, and a multi-stage test with four stages and a 1-3-4-5 design of modules) resulted in improved measurement precision as defined by RMSE and bias relative to paper-and-pencil testing. The quality of the measurement from those MST designs was almost as good as that observed with the CAT designs under study as well.

Studies by Kim (1993) and Kim and Plake (1993) also focused on two-stage testing. The purpose of the former study was to compare an IRT-based two-stage test to an individualized CAT to ascertain the conditions when two-stage testing might be an acceptably close alternative to CAT in terms of accuracy and efficiency of measurement. Variables of interest for the MST designs under consideration included differing the length of routing tests, the distributions of item difficulty parameters in the routing tests (peaked and rectangular), and number of modules in the second stage (6, 7, or 8). The results from this study indicated that a fixed-length CAT provided superior measurement precision for ability estimation to IRT-based two-stage tests of equivalent length. IRT-based two-stage tests using a rectangular distribution of item difficulty in the routing test and an odd number of second-stage tests produced more accurate ability estimates than did other two-stage test configurations studied.

In the Kim and Plake (1993) study, which was an extension of the Kim (1993) work, it was found that the structure and attributes of the routing test most substantially influenced measurement precision, but in most cases CAT was again providing more accurate ability estimates than any of the two-stage designs under consideration in this study. The best of the two-stage designs was the one with a rectangular distribution of

items in the routing test and an odd number of second-stage modules. In investigating the quality of measurement associated with two two-stage designs, Lam and Foong (1991) and Foong and Lam (1991) found that the recovery of true abilities for multi-stage tests was better than for comparable linear tests.

In Patsula (1999), 12 different MST designs were considered, also relative to CAT and paper-and-pencil. These designs varied with respect to the number of stages (2 or 3), the number of modules in each second and third stage test (either 3 or 5), and the number of items in each stage (between 6 and 24 in Stage 1, between 12 and 24 in Stage 2, and between 6 and 18 in Stage 3). As evaluated on the basis of RMSE, bias, and relative efficiency, the errors in ability estimation decreased as more stages and/or module per stage were added, though changes in the number of items per stage seemed to impact little on the quality of ability estimation.

However, for credentialing examinations, while individual proficiency estimates are important, the primary outcome of consequence is the classification of candidates into pass-fail categories on the basis of such scores. Thus, the second approach taken in studies of MST designs has focused more purposefully on the making of those binary pass-fail decisions using different test designs including MST. These results have generally been evaluated in terms of decision accuracy (DA) and decision consistency (DC).

DA indicates whether a decision made about a candidate reflects the truth, in that it is computed as a proportion of decisions that are consistent with the true decision classifications over all candidates. Similarly, DC reflects the consistency or stability of decisions for individual candidates made over parallel forms. If it were realistically

possible to have candidates take the same test twice or to administer parallel forms of a test to each candidate, DC uses proportion of decision agreement over replications to provide insight into reliability of the tests. Though such information is not obtainable in most live testing situations, in simulation research it is helpful for understanding the properties of the MST design given different conditions.

With respect to DA, to evaluate whether a given test design properly categorizes masters and non-masters, researchers in simulation studies compare true and observed classification decisions for Type I and Type II errors given a particular cut-score for making pass-fail decisions. The Kappa coefficient is also helpful in this task, in that it measures the agreement between the decision based on truth and on estimated ability, adjusted for agreement that might be expected to be due to chance factors alone.

Xing (2001) found that three CBT designs (linear parallel forms, MST, and CAT) provided essentially comparable results (as defined by DA, DC, and Kappa) in a simulation study investigating the effects of item quality, bank size, and placement of passing score. In this study, the passing score is moved but not to see where its placement might maximize DA and DC. Rather, the notion is that committees may want it set in different places. The placement of a passing score should be based on a consideration of content. Within each design, enhancing item quality and enlarging the item bank resulted in significant improvements in terms of the criteria of interest for pass-fail decision-making. In a follow-up study by Xing and Hambleton (2002), choice of test design was again found to be far less of a factor in terms of minimizing Type I and Type II classification errors than was bank size and quality. These authors suggest that when the pass-fail decisions are the primary objective of an examination, the complexity and

effort associated with adaptive test designs may not be entirely justified from a resource-allocation perspective: it may be as or more effective for test developers to administer a linear test and instead focus development on mechanisms for improving the item bank.

However, given that result, another study (Hambleton and Xing, 2002) was undertaken to explore the issues further: in that study, the focus was on optimal and non-optimal designs for linear parallel forms and MST. There, optimal and non-optimal is defined as relative to higher measurement precision in either the region of the cut-score for passing or in the region of the proficiency scale where many of the examinees are located. It was found the distinction made little practical significance, in that all of the designs investigated provided measurement results that were better than random selection, although in the case of the linear tests matching the test to the distribution of examinee proficiency did deliver slightly better results in terms of DA and DC.

In a recent study by Jodoin, Zenisky, and Hambleton (2002), a 60-item three-stage MST was compared with a 40-item two-stage test as well as several 60-item LOFT forms and the original, 60-item, operationally-used, linear test forms. The item bank used in this study was composed of item parameter estimates from 240 items from a paper-based section of a large-scale credentialing examination. While the results from all test designs were by and large comparable with respect to DC and DA, the three-stage MST and the LOFT forms provided results that were only minimally better than the original operational tests. This was in part due to the difficulty encountered by the ATA software in meeting the target information functions for the multi-stage and LOFT designs due to stringent content constraints. Interestingly, however, the results for the

two-stage MST (which, at 40 items, was two-thirds as long as the 60-item three-stage MST) were only very slightly lower than those observed for the three-stage MST.

Jodoin (2002) went on to explore these same designs (LOFT and two- and three-stage MST) with an item bank simulated to be improved with respect to both size and item quality. In addition, TIFs were varied with respect to information. What this study found was that as before, neither of the MST designs provided sizeable measurement advantages over linear forms, either in terms of correlations between true and estimated thetas or DA, although the longer MST and linear tests did result in slightly better DA.

2.4.2 Automated Test Assembly (ATA)

ATA is a particularly rich area for research into CBT in general and MST in particular. With respect to the state of ATA research, one particularly promising approach is the normalized weighted absolute deviations heuristic (NWADH, Luecht, Nungester & Hadadi, 1996; Luecht & Nungester, 1998; Luecht, 2000) that uses item-level information functions to manage need and availability of items in the bank to assemble modules and/or panels as specified by constraints. Other work by Armstrong, et al. (2000) and Reese, Schnipke, and Luebke (1999) has invoked a weighted deviations model in a process that involves the selection of items at random from the item bank to create modules. Berger's 1994 work on building optimal modules either within- or between-modules used test information in an item selection methodology predicated on estimating ability as efficiently as possible. This technique is, however, limited by the ability-level specific meaning of optimal, in that what is optimal for one ability level (range) is clearly not for a different level.

Van der Linden and Adema (1998) presented another method for ATA using linear programming (LP) where they conceptualize a multiple-form assembly problem instead as a series of two-form assemblies. LP was also the subject of an earlier study by Adema (1990) in which a variation on LP referred to as mixed integer linear programming (MILP). Such MILP models, as noted by Adema, are comprised of both integer and continuous decision variables. In this paper, Adema also modified a zero-one linear programming (termed ZOLP) approach for use in assembling an MST. Van der Linden (2000) presented several alternative methods for ATA based on mixed-integer programming for assembling tests from a bank with an item-set structure. These methods were evaluated using mathematical programming feasibility and expected solution times.

Luecht (1997), Vos (2000b), and Vos and Glas (2001) have also studied an aspect of ATA for MST that otherwise has not been studied previously: the case of building tests or modules with multidimensional constraints. As multidimensional IRT (MIRT) is increasingly being studied for eventual use in operational testing, its application to MST is a logical extension of previous research. As reported by Luecht, in the multidimensional case, TIFs are needed not only for total test or modules but also for separate content areas in which subscores are to be reported.

With so many such approaches to ATA, finding a methodology that aligns with the goals of different testing programs is possible. Ideally, however, with respect to MST, these automated test assembly algorithms not only need to be flexible enough to develop modules for various MST designs but also should be capable of creating multiple panels that control the overlap of items or modules between panels. For test development, such an approach can improve efficiency with respect to the basic assembly

of modules and permit great attention to be paid to those aspects of test assembly that are not so easily automated. There are qualitative concerns (for example, sensitivity and fairness issues) that are not so easily managed via automation, and those aspects of a test or module clearly benefit from careful review by test developers.

2.4.3 Special Applications of MST

Though the general MST model is of adaptive by stage testing, research into MST does include several specialized variations on the basic design. These lines of research, which include situations of non-adaptive multi-stage tests and modules where items are associated with a common stem, are described below.

2.4.3.1 Non-adaptive MST

At the outset of discussing the basic multi-stage model, MST was positioned in the middle ground between linear fixed-form and computerized-adaptive tests that are adaptive by item as a test design that affords test developers with some of the advantages from both CFT and CAT. Within the broad framework of MST, however, there is a substantial body of work relating to alternative MST structures for classification that select modules at random rather than based on previous module performance. In traditional MST, module selection is presumed to be adaptive in that modules are built to reflect several pre-specified difficulty levels and examinees are routed to them accordingly based on some routing system (be it IRT-based proficiency estimates or another approach such as number-correct scoring).

However, in computerized mastery testing (CMT), as described by Lewis and Sheehan (1988, 1990), the methodology is a testlet-based structure that implements variable-length multi-stage tests with more stages (and hence more items) being administered to candidates as needed to fulfill the chosen stopping rule (Folk & Smith, 2002). Candidates whose estimated ability places them far above or below the cut-score receive shorter tests, while those who are closer to the cut would be presented with additional sets of items as need to make a mastery decision. The goal in CMT is to minimize test length while simultaneously focusing on classification accuracy, and to do that all modules in CMT are constructed to be approximately equivalent to one another in terms of difficulty and content. Lewis and Sheehan (1988, 1990) used Bayesian decision theory and loss functions to minimize misclassification and test length. Because categorization into groups of masters and non-masters takes precedence over the quality of measurement along the ability scale, not all points on the scale are equally important in the basic CMT model.

After administration of a predetermined number of modules, examinees are either 1) categorized as masters or non-masters or 2) presented with additional modules if the desired level of precision relative to the cut-score is not met. If it is determined that testing should continue, this process of administering modules to an individual examinee keeps on until such precision is obtained or a maximum number of modules are presented. CMT is not, however, adaptive in terms of module selection, only with respect to the stopping rule. In the Lewis and Sheehan (1990) study, each examinee received between 2 and 6 testlets that were 10 items long (each testlet represented a stage). The testlets were constructed to be parallel using a variation of Lord's (1980)

procedure for fixed-length mastery testing, and given the decision to use random testlet selection the testlets were all peaked in difficulty around the score where examinees would be classified into master/non-master status.

Other studies of CMT include Sheehan and Lewis (1992) focusing on nonequivalent testlets and Smith and Lewis (1995) with multiple cut-scores, as well as Smith and Lewis (1998, 2002), Vos and Glas (2001), and Yi, Hanson, Widiatmo, and Harris (2001). Du, Lewis, and Pashley (1993) explored an application of fuzzy set decision theory for determining stopping rules rather than Bayesian approaches. In this approach, sets are defined by masters, nonmasters, and people for whom testing should continue: the sets are 'fuzzy' in that set membership is not according to some hard-and-fast rule as in traditional set theory but rather are a more relationally based on degree of set membership. They found that fuzzy set methods in a Rasch measurement context provided results that were quite comparable to the earlier study by Sheehan and Lewis (1990). In addition, research into additional other adaptive selection strategies in mastery testing has been developed (e.g., Kingsbury & Weiss, 1983; Reckase, 1983; Vos, 2000a, b), but by and large these algorithms are designed to work at the level of individual items rather than sets of items. With future research it may be possible to incorporate some of these methods for use with in adaptive MST.

2.4.3.2 Modules with Common-Stem Items

The second specialized area of MST research is focused on a particular module structure, specifically the case where the items within the module are not conditionally independent from one another (Wainer & Kiely, 1987; Wainer & Lewis, 1990). In that

situation, the appropriateness of some IRT models for adapting the modules to examinee ability is directly called into question.

By way of background, in IRT-based testing, there are two important and related assumptions about test items that underlie that use of IRT. The first of these is that individual items are locally independent from one another, and the second key assumption is of unidimensionality. These assumptions are related because in the case where the local independence assumption is violated, something other than examinee ability is influencing responses. That something is a dependence between responses to individual items that changes the dimensionality of the test form.

Such dependence is a problem in the context of IRT-based MST where the modules are composed of sets of items linked in some way such as a passage or graphic, because research has demonstrated that in such cases reliability of the test composed of such sets of items tends to be overestimated, resulting in overconfidence in the precision of examinee scores (Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Zenisky, Hambleton, & Sireci, 2002). In addition to problems with estimation of reliability, item sets based on a common stem have also been investigated for the presence of differential testlet functioning (a generalization on studies of differential item functioning; see Wainer, 1995; Wainer, Sireci, & Thissen, 1991). Lee and Frisbie (1999) also developed an approach to estimating the reliability of such modules using generalizability theory.

In dealing with such testlets with respect to estimating examinee scores, the common approach has involved scoring methods using polytomous IRT models (e.g., Thissen, Steinberg, & Mooney, 1989), although such an approach was not appropriate for adaptive testing. While polytomous models may be useful in that conditional

independence between the item sets can be retained, the use of polytomous models also results in a net loss of item information because not all parameters are estimated for each dichotomously-scored item within the polytomous item set. For example, with the graded response model of Samejima (1969) a single discrimination parameter for the polytomous item is computed, along with a threshold value for each score point.

However, recent research efforts have been directed toward alternative methods for conceptualizing and analyzing modules with items linked in this manner that do facilitate adaptive testing. This is an important emerging area of research for MST. Work by Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (2000) in what has come to be described as testlet response theory has brought about the development of modifications to the two- and three-parameter logistic IRT models which allow for on-the-fly construction of item sets that appropriately meet constraints including the minimization of local dependence. In the former case, the model actually includes an extra parameter to represent the interaction effect between an examinee and a given testlet, while the second study is a further generalization of the previous work, but due to added complexity in the 3PL model this methodology is more intensive computationally. Further work in this regard has also been done by Vos and Glas (2000) and Glas, Wainer, and Bradlow (2000).

One recent study by Rotou, Patsula, Steffen, and Rizavi (2003) explored the case of using set-based items with MST comparing the results to the tests where set-based items were administered as a) CAT and as b) paper-and-pencil nonadaptive tests. In this study, the researchers had access to 440 items (which comprised 64 item sets) testing verbal reasoning ability. The first comparison entailed simulation of a 32-item CAT test

and a 33-item two-stage MST (16 items in Stage 1 and 17 items in each of three Stage 2 modules), and then results from simulated administrations of a 55-item P&P test and a 54-item two-stage MST (23 items in routing stage and 31 items in each of three Stage 2 modules) were similarly compared. The outcomes of interest for this study were focused on indices of measurement precision, also as related to the choice of the 1-, 2-, or 3-parameter logistic IRT model for calibration and scoring. Results indicated that MST provided better results than the equivalent-length P&P test, and the results for the CAT were also of the same level of reliability as the P&P test. With respect to choice of IRT models, the MST design actually gave slightly better results in terms of reliability with the 1- and 2-PL than the CAT, while with the 3-PL CAT and MST were largely equivalent.

2.5 Summary

As the stakes associated with much educational and psychological testing continue to increase, more and more attention is being paid to issues such as the role of measurement errors and misclassification. For testing programs, particularly in the area of certification and licensure (where agencies have the dual responsibilities of providing fairness for candidates and protecting the public), obtaining highly precise measurement and decision accuracy are critical part of establishing test score validity. This is particularly the case in CBT applications such as MST and CAT where technologies for administration and test development are changing and being updated with incredible speed. In that regard, the goal of trying to identify the single 'best' approach or design structure in MST for practice even within the general domain of testing for certification

and licensure is not a practically viable one. However, efforts to ascertain general psychometric properties associated with various design variables of an MST can be useful as agencies interested in the use of MST go about the process of designing feasibility studies and assessing the costs and benefits (both measurement and otherwise) for their testing programs associated with instituting a computer-based multi-stage test.

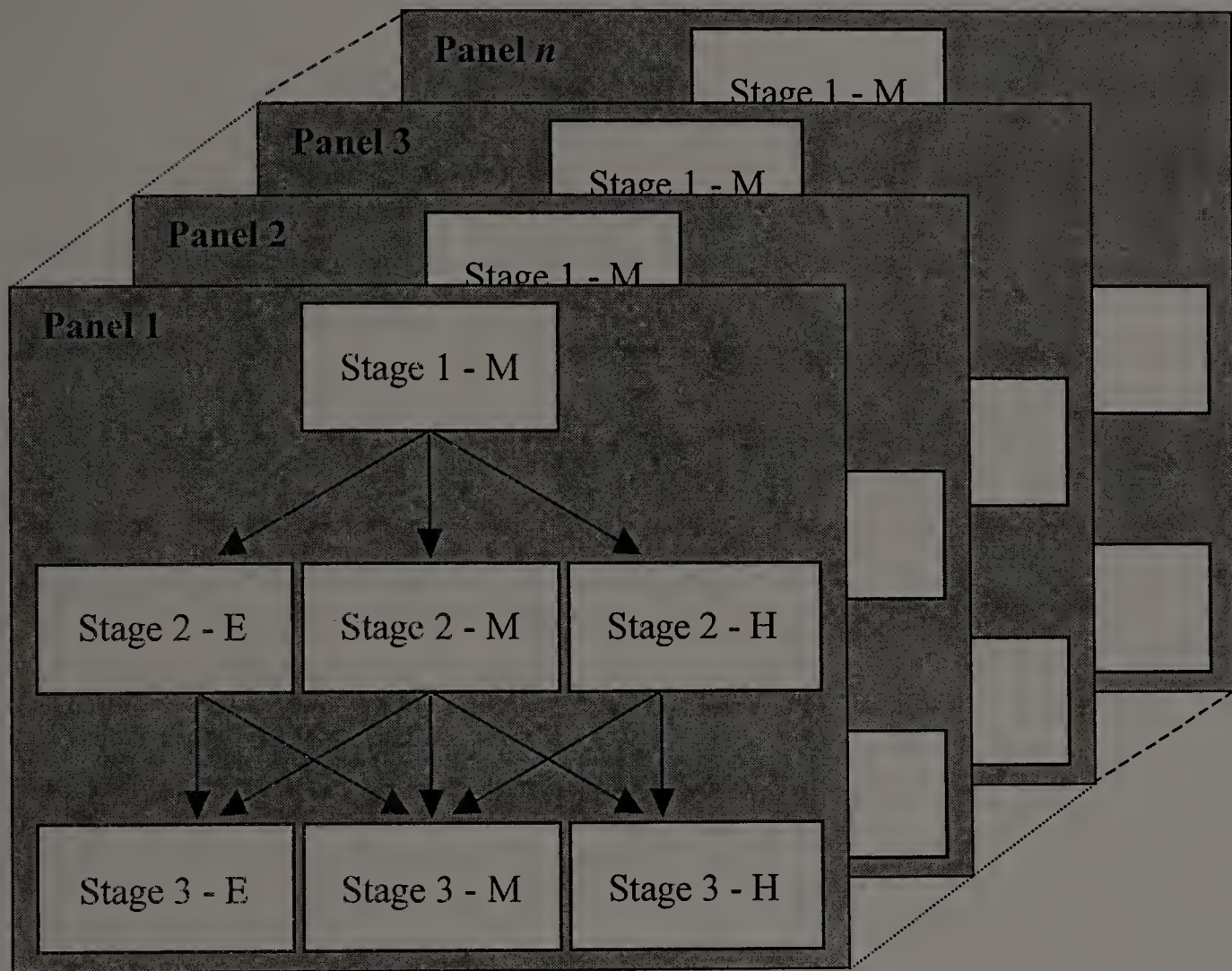
For professional credentialing assessment, multi-stage testing can be viewed as an effort to capitalize on the efficiency of CAT and the test form assembly controls of linear testing. Through this review of the MST literature, it is particularly clear that the relative benefits of MST are very much dependent on the characteristics, needs, and goals of individual testing programs. Issues such as (but not limited to) depth and breadth of the item bank, the selection of automated test assembly algorithms, the specific design structure implemented, and the placement of the cut-scores for making the critical pass-fail decisions are just a few of the essential variables that must be deliberated upon the process of developing such a test.

Among these variables, several have emerged as potentially having a great deal of practical significance on results for candidates. The choice of design, the amount and distribution of test information, and the test length are all variables with such promise. In addition, routing methodologies are an important and relatively understudied aspect of MST. To date, the focus of MST research has been toward the 'front-end' of development, specifically toward the more structural variables and the test development aspects. Given that MST is not a test design widely used as yet operationally, attention to this aspect of the approach can be understood as the next logical direction for research attention. Only a relatively few strategies have been tried, including routings based on

number-correct scoring and population distributions, and the literature does not seem to reflect any studies that have empirically compared any of the proposed strategies for either accuracy of ability estimation or in the context of classification. While the methods used presently seem to work sufficiently, it seems clear that the measurement effectiveness of the design is predicated on the nature and defensibility of the routing decisions and as such it is only with additional research efforts in this design aspect that high-stakes decisions can be made on the basis of scores from a multi-stage test.

To this end, it seems clear that work is cut out for continued research into MST. A multi-stage test is a highly complex and variable test design, but as noted previously, such variability can be viewed as an advantage in terms of design flexibility. A multi-stage test can be built to greater resemble a CFT or a CAT: ultimately, the design strikes a balance of adaptability, practicality, measurement accuracy, and control over test forms. With the current MST literature and the study proposed here, as the relational effects between different design variables are delineated, the potential exists for MST to take on an increasingly significant role as a viable alternative for more and more agencies involved with the important task of assessment for professional certification and licensure.

Figure 2.1. Illustration of Parallel Panel Structure



CHAPTER 3

METHODOLOGY

3.1 Overview of the Study

The purposes and methodology for the present study are described in this chapter, including an overview of the variables and conditions implemented. In the course of detailing the methodology, a step-by-step explanation of simulating a multi-stage test is also provided. In this regard, there are two primary components. The first of these was automated assembly of modules and forms using a specified item bank and test information function. Second, after assembly, the actual simulation of examinees taking the modules and forms was completed using a simulation program, which provides ability estimates and other results that are used for data analysis.

The study was designed as a study of selected variables for implementing MST, and is intended to further develop previous research into the psychometric properties of multi-stage testing (MST) for certification and licensure, with the primary goal being to extend understanding of how these variables affect test scores and decisions about candidates. The primary design variables of interest in this study were 1) test structure, 2) the amount of test information, 3) the distribution of test information across stages, and 4) different between-stage routing strategies. These four variables were considered in the context of three passing rates (30%, 40%, and 50%), chosen for providing a range for interpreting results and due to relevance to the agency whose data was modeled in this study. The rationale for including these variables in this study is likewise provided in some depth below.

To start, one important aspect of this study involves how several MST test structures (specifically the 1-2-2, 1-2-3, 1-3-2, and 1-3-3 combinations of modules and stages) compared relative to one another with respect to ability estimation and decision classification. While a great deal of the MST research has looked into the 1-3-3 structure, few studies to this point have provided feedback on what levels of accuracy might be expected given other designs, and the 1-2-2 approach is particularly understudied in this regard.

Second, based on the previous studies of MST (Jodoin, Zenisky, & Hambleton, 2002; Jodoin, 2002; Xing & Hambleton, 2002; Zenisky, 2002), a critical need in the MST literature concerns further investigation of the impact of varying amounts of test information. In previous research, the level of test information specified was in most cases defined by the test information functions used in operational testing, and this resulted in high rates of decision accuracy and consistency. A reasonable direction for further research concerning this aspect of the design is to try and to vary the amount of test information called for in the test information function (with regard to both increases and decreases). Implementing these sorts of variations may provide insight into the levels of decision accuracy and consistency that can be expected with such target TIFs. How much practical benefit does increasing the target TIF by 50% bring, and likewise, what does proportional reductions of the target TIF mean for ability estimation and decision classification? This is a variable of interest to test developers working with the MST design because with lower levels of target test information, the test assembly mechanism has the potential for greater flexibility in putting together test forms with regard to the statistical characteristics of an item bank. If the algorithm can draw on

items with more varied discrimination and difficulty values in order to meet its target TIF, it may well translate into greater use of the item bank in terms of both breadth and depth, thereby reducing item exposure. In this way, if the ability estimation and decision classification results are of somewhat comparable quality to the results obtained with current levels of information, test security can be enhanced while still providing high-quality assessment results in terms of making pass-fail decisions.

Third, additional study into how test information should be distributed across stages is warranted. In previous studies, such target information has generally been split equally among the three stages with satisfactory results, but previous research (c.f., Zenisky, 2002; Jodoin, Zenisky, and Hambleton, 2002) indicates that alternative distributions such as 1/2-1/4-1/4 may provide improved decision accuracy and consistency because elevated levels of measurement information are collected early in the test. By obtaining better ability estimates after the first stage, it may be possible to make routing to second- and third stage tests more efficient, thereby improving score precision at the conclusion of testing for many candidates. It seems clear that further research into this aspect of MST can substantially clarify how this variable impacts the quality of measurement obtained.

The fourth goal of this study is to evaluate routing strategies in the context of several MST designs. Several different methods for routing can be found in the psychometric literature, although such methods have not yet been empirically compared by means of a simulation study. A description of the four methods to be implemented in this study can be found in Section 3.5.1. This is a variable of importance because the choice of method used for routing examinees between stages is fundamental to the

process of adapting a multi-stage test to candidate ability, and empirical determination of whether certain strategies provide more optimal results than others can help to ensure that the test design implemented is as efficient as it can be and provides test results that are psychometrically accurate to a very high degree.

The last goal of this study was to consider all results in the context of three different passing rates. By varying passing rates on a percentage-passing scale (30%, 40%, and 50% passing), the generalizability of this study can be enhanced. These values are chosen to reflect passing rates that are seen in operational testing with programmatic variables similar to those simulated here. Though testing programs do not generally set cut-scores using norm-referenced criteria, the effect of using of different passing rates as operationalized in this study is to permit analysis of decision classification outcomes at three different places on the ability scale where a criterion-referenced cut-score might be placed. As the cut-score moves from more (50%) to fewer (30%) passing, of interest was how this might impact on the probability of misclassification for candidates given the other variables under consideration in this study.

In total, there are 384 conditions to be evaluated in this study (4 levels of total test information by 2 levels of the distribution of test information across stages by 4 levels of test designs by 4 levels of routing strategies by 3 levels of passing rates). The next sections detail the methodology for this simulation study, including a focus on the generation of the item bank, the creation of modules and tests for the conditions to be simulated in this study using the computer program CASTISEL (Luecht, 1998) for automated test assembly, the rationale for selecting such conditions, the simulation of the

multi-stage tests with MSTSIM5 (Jodoin, 2003), and the procedures taken in terms of data analysis.

3.2 Simulation of the Item Bank

A secondary finding of recent research by Jodoin, Zenisky, and Hambleton (2002) was that it was difficult to adequately meet target test information functions (TIFs) with a bank consisting of 238 operational Financial Accounting and Reporting (FAR) items calibrated using the three-parameter logistic model of Birnbaum (Lord & Novick, 1968) from recent paper-and-pencil forms of the American Institute for Certified Public Accountants' (AICPA) Uniform CPA Exam. This finding was not unexpected, but indicated the need to expand the item bank in any subsequent research involving modeling of the AICPA's tests. Even with an additional two forms' worth of items (for a total of 358 items), the possibility of confounding the issues of limitations of the item bank and measurement information led to further reservations about using only these 358 items for this study. For this reason, a larger item bank was generated for this study.

In generating the item bank, the primary consideration was to develop and use a bank that reflected the kind of bank that would be seen in operational testing by an agency such as the AICPA. By building a larger item bank to reflect the statistical realities of the bank of 358 FAR items, it was determined that two particular statistical dimensions needed to be reflected in any simulated bank. The first of these was the means and standard deviations of the item parameter estimates and the second was the correlational structure among these estimates. Preliminary analyses suggested that a bank size of 2,500-3,000 items would provide sufficient breadth and depth for building tests

with the desired levels of difficulty and discrimination. Though this bank size is considerable, the target information functions being specified are not unusually high for a high-stakes test of the nature being simulated. The bank and its size were produced to ensure that the targets could be met to the greatest extent possible so that any results obtained would not be confounded by the TIFs falling short of the desired levels; hence the large item bank.

To build this bank, a statistical technique to 'clone' the 358 current items was employed. This approach involved adding or subtracting a small error term (determined by a random number generated by a uniform distribution between 0 and 1) to each item's three parameters (discrimination, difficulty, and guessing) in the current bank. The new parameters created through this approach closely approximate the means and standard deviations of the parameter estimates in the current bank as well as the inter-correlations between the items. This technique was repeated on the 358 items in the original bank 7 times, thereby creating a bank totaling 3,222 items (358 original items and the 2,864 'cloned' items). Descriptive statistics for the original set of items and the item bank generated in this way were identical across the original and generated item banks, with the means (and standard deviations) for the a -, b -, and c -parameters constant at 0.62 (0.25), -0.12 (1.11), and 0.00 (0.30). In the original set of 354 items, the correlation between the a - and b -parameter estimates was 0.363, between the a - and c -parameters it was 0.346, and between the b - and c -parameters the correlation was 0.308; among the items in the generated bank, the correlations were 0.362, 0.337, and 0.298, respectively. These statistical parallels ensured that the generated bank was a reasonable reflection of the current, operational bank, only larger.

3.3 MST Test Structures

A test structure for MST was defined as a particular combination of modules and stages that together comprised all of the potential routes that an examinee could take in the course of being administered a multi-stage test. For example, in the 1-3-3 structure, a candidate would at first be given a medium-difficulty module, and in both of the subsequent stages could receive an easy, medium, or hard module. Module difficulty was defined by the average of the statistical characteristics of the items that comprised the module, and was specified in test assembly by the positioning of the module-level information functions. The more 'to the left' a module information function is located on the ability scale, the easier the module, and similarly the more 'to the right', the harder the module. On average, the items in an easy module are easier than those in a medium difficulty module, and items that are harder on average will be included in harder modules. Easy, medium, and hard are of course relative terms, and are given meaning by the positioning of the target module information function relative to the ability scale.

Four specific test structures were of interest in this study: the 1-2-2, 1-3-3, 1-2-3, and the 1-3-2 (Figure 3.1). Across all four designs, the target information function for the first stage module was always targeted at the passing score, regardless of whether the condition is 30%, 40%, or 50% passing, and the modules in subsequent stages were positioned relative to the passing score as well.

In the 1-2-2, design, all examinees were presented with a module of medium difficulty in Stage 1. There were two modules to which examinees can be routed to in the second and third stages respectively, either relatively easier (for weaker examinees) or relatively harder (for more able examinees). A similar structure was implemented in the

case of the 1-3-3 MST design. There, examinees were again first given a module of medium difficulty in Stage 1, and depending on routing strategy were routed to either easier, medium, or harder modules in Stages 2 and 3.

The 1-2-3 and the 1-3-2 designs can be represented similarly. In the former case, all examinees receive a middle difficulty module in Stage 1, at which point examinees are routed to one of two modules in Stage 2, either an easier or a harder module. From there, examinees are sent to either an easier, medium, or harder module in Stage 3. The 1-3-2 structure starts with a medium-difficulty module in Stage 1, and routes examinees to easier, medium, or harder modules in Stage 2. For Stage 3, the test structure is set up such that examinees receive either an easier or harder module, depending on their estimated ability.

3.4 ATA using CASTISEL

The computer program CASTISEL (Luecht, 1998) was used for ATA in the present study. CASTISEL is an automated test assembly program that takes statistical and other content constraints into account and automates the process of formulating modules and panels for MST using the item bank specified. With CASTISEL, MST modules were simulated by specifying target information functions, the number of stages to be included in a form or module, the number of modules per stage, the number of items per stage, the total test length, and the primary content specifications for content balancing being implemented. CASTISEL creates such forms or modules by using the normalized weighted absolute deviations heuristic (NWADH) described by Luecht

(1998) to optimize item selection for forms or modules given the target TIFs and other form or module-level considerations.

By using a pre-calibrated real or simulated item pool, the NWAD heuristic as implemented in CASTISEL constructed one or more parallel test forms as needed by sequentially completing locally optimal searches to find items that meet the statistical and content constraints defined in the input file. In this context, test forms can be taken to refer to linear forms or to a set of modules as is needed for MST.

The NWAD heuristic can be understood in more detail as an algorithm based on several component parts (Boughton & Gierl, 2000). Normalization refers to the way in which various objective functions are selected in order that they can be met simultaneously, and the weighting involves prioritizing certain items with 'poorer' item statistics within content areas which allows for balancing between more and less discriminating items to take place. 'Poorer' items, in this context, were those items that in and of themselves may not have met certain specified minimum statistical constraints, but with this weighting, they could still be used in test assembly by taking those items into account along with items that exceed such constraints. The absolute deviation portion of the process describes the absolute difference between the target test information function and the current function. Lastly, a heuristic is an approach to problem-solving that iteratively evaluates and chooses the best-fitting answer or solution at specified points in the course of test assembly.

In the case of a multi-stage test, CASTISEL assembled as many sets of items (modules) as needed for each stage, each containing a unique sequence of items designed and selected to reflect certain content and statistical specifications. For example, with a

three-stage test with three 20-item modules in each stage (three parallel medium difficulty modules in Stage 1 and an easy, medium, and hard module in Stages 2 and 3), CASTISEL builds nine modules (here, three per stage), and each of those output files would contain 20 unique items selected to meet both the statistical criteria of target test information functions and content balancing based on a primary content dimension.

The specification of the target test information and other related aspects of modules were critical for automated test assembly. In the next sections are detailed the considerations that were taken into account with respect to this aspect of test assembly in this study.

3.4.1 Obtaining the Base Target TIF

The base target information function for this study was obtained via the TIF from operational tests: in this case, the average TIF from six archival forms of the American Institute of Certified Public Accountants' "Financial Accounting and Reporting" (FAR) test section was computed. These forms consisted of approximately 60 items each (2 tests had had one item deleted). Each of these forms was calibrated using the three-parameter logistic model, and those item statistics were then used to obtain TIFs for each of those forms. The average across those six forms can be taken as a reasonable representation of an average TIF as would be observed on this section of the AICPA examination. The six operational TIFs as well as the average TIF obtained are given in Figure 3.2.

3.4.2 Targeting Test Information to the Passing Score

The placement of the target test information function relative to the ability scale defined difficulty of the overall test: as an information function is moved to the left a test on average becomes easier, while shifting the information function to the right produces a harder test. Where to center the test was not an automatic decision: generally speaking, test information could be set to be optimal at either the passing score or where the majority of candidates were. In the latter case, for example, if there were a normal (0,1) distribution of examinees, then information could be centered at 0.0 to ensure maximal information for people in the middle of the ability distribution. The alternative, targeting test information to the passing score, means that score precision was maximized for candidates whose estimated ability places them in the vicinity of the cut-score for making pass-fail decisions, which may or may not be mapped to 0.0 on the ability scale.

In credentialing, decision classification is the primary outcome of interest: accurate classification for as many candidates as possible is critical. For very able and very weak candidates, the decisions are likely to be correct in the majority of cases. However, it is candidates who are near the passing score for whom the risk of misclassification is the highest, and for that reason in this study the tests (as composed of stages and modules) were constructed to be optimally informative at the passing scores. In this study, passing rates of 30%, 40%, and 50% were of interest, which translated to values of 0.521, 0.223, and 0.000 with respect to a normal distribution centered at 0.0.

The average target information function that this study was based on was centered at approximately 0.75 on the ability scale. Once this average target information function was obtained, it is necessary to re-align this information function to center at each of the

three pass rates. In Figure 3.3 are given the original and the three re-centered target TIFs used in this study.

3.4.3 Amount of Target Test Information

As noted previously, the total amount of test information used in this study was based on the average TIF for operational forms of a section of the AICPA's Uniform CPA Exam. To the extent that such information levels could be made generalizable to other credentialing agencies, it was of interest to try and vary the relative amount of test information at different ability levels. In reducing the total amount of test information, to what extent was ability estimation and decisions classification impacted by less test information? If certain levels of reduction in test information resulted in comparable measurement to the case where full information is used, then it may be possible to go with lesser amounts of test information or reduce the number of items in the test. An argument for reducing test information is to make the target test information functions somewhat less stringent, thereby freeing the test assembly software to be more flexible in meeting the targets given the statistics of the items in the item bank. Alternatively, increasing test information provides additional context for the levels of decision accuracy observed with lesser amounts of test information.

In addition to using the average target TIF at its current level of information, this target TIF was increased by 50% and reduced by 25% and 50% for the current study in order to evaluate the impact of varying target information on the outcome of interest to credentialing agencies. In Figure 3.4 the target test information functions at each amount of test information for each of the three pass rates of interest in this study are given.

These percentage reductions for the target TIF were selected as exploratory values by which test developers might be interested in either increasing or reducing information. The results of this analysis were intended to provide insight into the interaction between information levels and measurement accuracy.

3.4.4 Specifying Stage-Level Information

CASTISEL required that target TIFs be specified for each module in each panel to be assembled. The first step in doing this involved partitioning of the target test information function to create stage-level information functions. These stage-level information functions provided the statistical information needed by the automated test assembly software to build modules to achieve a particular level of measurement data about individual candidates.

In the general case, test developers have to decide what proportion of test information they wish to obtain in which stage, and portion out the overall test information function accordingly. One approach to this is to choose to obtain equal levels of information across stages, so where a three-stage multi-stage test is to be used, the overall target test information function is divided by three in order to provide an individual target level of information for each of the three stages. In aggregate, the stage-level information functions provide the appropriate level of test information desired after administration of all three stages. However, though equal information among stages does possess certain intuitive appeal, is only one of several possible strategies. Preliminary empirical results suggested that the use of alternative distributions, particularly methods

that provide higher test information early in the test, may additionally enhance the quality of routing decisions and, in the end, pass-fail decisions.

Thus, of interest in this study were two different distributions of target information across each of three stages in an MST: the first was equal information across stages and the second was the case where $1/2$ of the test information was obtained in Stage 1 and $1/4$ information was obtained in Stages 2 and 3 respectively. These two different partitionings of the target TIF were thus involved in this study, based on results from previous studies of MST using AICPA data (Jodoin, Zenisky, and Hambleton, 2002; Zenisky, 2002).

To accomplish these partitionings, the average TIF obtained from the six operational TIFs (as described previously) was divided up as needed. These two variations are specified as follows:

- To create a design with equivalent information in each of three stages ($1/3$ - $1/3$ - $1/3$), the average target TIF was divided by three;
- To create a design with $1/2$ of test information in Stage 1 and $1/4$ information in stages 2 and 3, the average TIF from above was divided by $1/2$ for the Stage 1 modules and quartered for the Stage 2 and 3 modules.

These distributions of target test information were selected to inform practice about not only specific proportions of information being allotted to different stages, but also to provide additional insight into how increased information at different between-stage points in the test compared to the case where equal information across stages was specified in terms of the quality of ability estimation.

3.4.5 Quantifying Within-Stage Module Difficulty Differences

In CASTISEL it was also necessary to use the stage-level target information functions to define how modules vary on the basis of difficulty. In the first stage of an MST, examinees are commonly presented with a module of medium difficulty, while in later stages there may be relatively easier and harder modules in addition to medium difficulty modules.

As described in Section 3.4.1.4, the target TIFs were to be centered at 0.521, 0.223, and 0.000 in this study. Thus, for medium-difficulty modules in any stage where the passing rate is to be 50%, the maximum value of the TIF was centered at that passing score (0.000). From there, relatively easy and hard modules can be specified using the stage-level TIFs translated to the right and or left by a certain quantity, such as 1/2 of a standard deviation. Alternatively, where the 30% pass rate was implemented, the target TIF was centered at 0.521, and the easy and hard modules were shifted to the left and right as appropriate.

The decision of how different to make the modules within each stage is one that in operational testing is dependent on two basic factors, with the primary goal being to create modules that are appropriately targeted to candidates of different abilities. These are 1) the distribution of the candidate population (how dispersed or clustered together are the bulk of candidates) and 2) the depth and breadth of the item bank (to support building modules of different difficulty levels).

For example, in a stage with three modules (relative easy, medium, and relatively hard) such as is found with the 1-3-3 MST design, the information function for the module of relatively easier difficulty can be specified by taking the stage-level TIF

associated with that stage and shifting it perhaps $1/2$ standard deviation to the left of where it is placed for a module of medium difficulty. Similarly, for a relatively harder module, the partitioned TIF is shifted perhaps $1/2$ standard deviation to the right of the TIF for the medium-difficulty module.

Given in Figure 3.5 is a visual example of what it means for such stage-level test information functions to correspond to modules that vary by difficulty. The dashed lines represent stage-level information functions, and (as is evident in Figure 3.5) are provided for every module to be assembled by CASTISEL. The example provided in Figure 3.5 is shown as just one of countless possibilities: of course test developers can specify different numbers of modules within stages, different numbers of stages, and more or less overlap of module difficulty as desired.

In the context of this study, modules varied from one another by $1/2$ standard deviation. This level of difference seemed reasonable based on two considerations. First, the study implemented a candidate ability distribution that is normally distributed with a mean of 0 and a standard deviation of 1, and second, exploratory analyses using different item banks based on the item statistics to be implemented here supported construction of this level of between-module difficulty differences (Zenisky, 2002).

3.4.6 Content-Balancing Test Forms and Modules

The last important consideration in ATA is content balancing. Content balancing, in this study and in many applications of MST, involves representation of content not just across the entire test, but also within each stage. Thus, if the test specifications call for nine items of a particular content area to be included across all three stages of a multi-stage

test, then three of those items would appear on stages 1, a different three in Stage 2, and yet another three items in for Stage 3. This balancing of content is consistent regardless of how an examinee is routed through the test: no matter what sequence of easier, medium, and/or harder modules an examinee sees, the number of items from a given content area is constant.

In terms of content-balancing, from the original six operational forms in the 358-item bank, average content specifications for the primary content dimension were obtained. There are three primary content dimensions (called 1, 2, and 3, for convenience). From there, using the proportions identified in the original test forms and module lengths of 20 items, it was possible to calculate the number of items from each content dimension that should be represented on each module in each stage of a multi-stage test (Table 3.1).

Thus, for example, in the case of the 60-item test, eight total items from Dimension 1 are needed across all three stages, and so three items from Dimension 1 would be included in Stages 1 and 2, and then two items from that dimension would appear in all second-stage modules as well. Similarly, whereas 22 items are needed from Dimension 2 for the entire test, that could be achieved by integrating seven Dimension 2 items in Stage 1, followed by eight Dimension 2 items in Stage 2 and seven Dimension 2 items in Stage 3. In each stage, 10 items from Dimension 3 would be included.

3.4.7 Assembling Parallel Panels

Using the specified item banks generated for this study, along with the different levels of test length and distribution of target test information, automated assembly of

modules and panels for each of the two test structures under consideration in this study was then completed.

For the 1-2-2 design, for each module shown in Figure 3.4, two identical stage-level target information functions were specified. In terms of MST this meant that CASTISEL constructed two parallel panels, each consisting of five modules in the arrangement specified in Figure 3.1. This results in 10 total modules being built for each condition with the 1-2-2 design (where conditions are defined as combinations of the variables of interest: item bank size, test length, passing rate, and distribution of test information). In this simulation, exposure of each second and third stage module could thus be controlled to 25%, while Stage 1 modules are exposed at 50%. To clarify: for the 1-2-2 design, all candidates routed to Panel 1 see one common module in Stage 1, while all candidates receiving Panel 2 see a medium difficulty module parallel in difficulty to what is seen by candidates in Panel 1. From there, in both panels, examinees were routed to second and third stage easier or harder modules as determined by ability estimates. The easy modules in the second and third stages were constructed to be parallel to one another both within and across the two panels, and this also applies for the harder modules.

For the 1-3-3 design, three parallel modules of the Stage 1 medium were built, and one each of the modules in Stages 2 and 3 were built. Thus a total of nine modules for this design were constructed by CASTISEL, and in this case, module exposure across all three stages is limited to 33%. In effect, in terms of the 1-3-3 design, three parallel panels are used. These panels only differed from one another by the first stage module, as three parallel medium difficulty modules were created to maintain item and module

exposure for the first stage at 33%. The modules in the second and third stages of the three panels were identical. For the 1-3-2 and 1-2-3 designs, three parallel panels of each involving different first-stage modules for each panel will be assembled.

3.4.8 Meeting Target TIFS

A critical aspect of IRT-based test construction is the concept of test information. As noted by Hambleton, Swaminathan, and Rogers (1991), the amount of test information at each point on the theta scale can be understood as relating to the level of measurement precision that could be expected at that point. Thus, when the information value at a point on the ability scale is high, that translates into highly precise measurement: conversely, lower information means a decrease in precision. By specifying target information functions for each module, and tailoring test information to different ability levels as is done with MST, it is possible to improve measurement precision at more points along the ability scale than can otherwise be done with a single TIF on a linear fixed-form.

Implementing such an approach was predicated on being able to successfully meet target TIFs. Thus, in terms of test assembly using the NWADH (as implemented in CASTISEL), it is important to consider the extent to which the item bank can support the building of modules and panels to the prescribed test information specifications. In this portion of the analyses, mean TIF differences and mean square errors (MSE) of the TIF difference were computed by CASTISEL for each module assembled. The mean TIF difference was the averages of the deviations for the items in a module across a range of θ values from -2.0 to 2.0 , and MSE for the TIF differences is the average squared

deviation for each module. MSE values for modules were then averaged across modules within an analysis condition to obtain a MSE index for each analysis condition, and these values were inspected for magnitude. By using this statistic, it is possible to evaluate an index of a rough but immediate indication of the extent to which targets were met in the context of the simulation.

In this study, MSEs of the TIF differences for the various conditions ranged from .00 to .18, and mean TIF differences similarly ranged from .00 to -.08. These values across conditions in this study were in large part consistent with those reported by Patsula (1999). Thus, these results suggest that the process of selecting items to match the statistical information specified by the target TIFs in this study produced operational TIFs that were practically accurate for the purposes of simulation.

3.5 Simulating a Multi-Stage Test

The modules assembled by CASTISEL in each of the simulation conditions were then used to simulate MST administration by means of the computer program MSTSIM5 (Jodoin, 2003). MSTSIM5 was designed to use the output of CASTISEL in simulating examinees and test designs. As input, MSTSIM5 read in the item sequence information for each module in a MST panel contained within the CASTISEL output files, and simulated examinee responses to items within modules and panels as assembled by CASTISEL. Sample size was set as desired, and in MSTSIM5, it was also possible to specify the distribution of simulated candidates. Ability estimation in MSTSIM5 is done by means of maximum likelihood estimation.

To determine the distribution of candidates for this study, analyses of the ability distribution of AICPA candidates were completed using examinee response data from archival administrations of the FAR section of the Uniform CPA Exam. An inspection of the results revealed that, for multiple forms, the underlying distribution of candidate ability was normal with a mean of about 0.0 and a standard deviation of about 1.0. Thus, to be consistent with what is observed in operational testing by an agency such as the AICPA, the distribution of candidates for this study is set at normal with a mean of 0.0 and standard deviation 1.0. Sample size for this study is 9,000 examinees, chosen to eliminate sampling errors.

3.5.1 MST Simulation with Varying Strategies for Between-Stage Routing

The MSTSIM5 program (Jodoin, 2003) was used to simulate the multi-stage test administration. However, as routing strategies implemented in MSTSIM5 were a variable of interest in this study, the original MSTSIM5 program was modified to implement the four routing methods described below.

3.5.1.1 Defined Population Intervals (DPI)

This approach defined relative proportions in population expected to follow each of primary routes, and routing within each stage of the MST occurs based according to such proportional assignment. This method has likewise been widely used in recent studies of MST (i.e., Jodoin, Zenisky, and Hambleton, 2002; Jodoin, 2002; Xing, 2001; Xing and Hambleton, 2001; Hambleton and Xing, 2002). With this strategy, for routing to the second and further modules, assignment of modules was done based on previous

ability estimates. With a normal distribution of candidates (from a population with mean 0 and standard deviation of 1), testing programs can empirically determine the cut-scores for routing to ensure certain levels of module exposure, such as equal assignment of candidates to modules. This method was relatively simple to implement and allowed for exposure rates to be known in advance of testing, but it may be inappropriate to route candidates through a test structure to make a criterion-referenced decision using a norm-referenced methodology.

For example, it was possible with this approach to ensure that the lowest one-third of candidates was directed to the easy module, the highest third to the hard module, and the middle third could be sent to a module of intermediate difficulty. Ensuring the equivalence of assignment across modules was the apportioning strategy used to set the cut-scores for this approach in this study. In conditions where there were three modules in an impending stage (as in the 1-3-3 and at points in the 1-2-3 and the 1-3-2 designs), the simulee population was rank-ordered by provisional theta estimates from the stage immediately previous and split into thirds for assignment to modules in that upcoming stage. Given a normal distribution of ability in the simulee population, the two cut-scores needed for a stage with three modules were set at -0.43 and 0.43, reflecting the points on a normal curve that divide the area under the curve into thirds. Where there were two modules to a stage (as is the case for the 1-2-2 and at points in the 1-2-3, and 1-3-2) the simulee population was similarly rank-ordered but divided into lower and upper halves based on provisional ability estimates, with the cut set at 0.0.

3.5.1.2 Matching Module Difficulty and Ability Estimates (Proximity)

This approach was used by Kim (1993) and Kim and Plake (1993), and involved assignment of n -stage modules that varied in average difficulty to candidates based on a proximity calculation. In this approach, the average difficulties of each module in a given stage were computed, and the most recent provisional theta estimate of each candidate was compared to those average difficulties. The candidate was routed to the module in that stage for which the difference between the module difficulty and provisional theta estimate was the smallest, thereby providing a mechanism for assigning candidates to modules at the (approximate) appropriate difficulty level. This comparison was repeated at each juncture between stages where assignment of modules in the next stage was required. By routing examinees in this way, it is important to note that the distribution of candidates across modules within stages 2 and 3 (regardless of design) would vary somewhat, and consequently some modules would likely be more exposed than others (as compared to some other routing strategies that might be used in which ensuring equal assignment of candidates to modules is emphasized, such as the DPI method).

One advantage associated with this approach to routing candidates through an MST was that it was relatively simple to implement, although such a method may be problematic for operational use due to item and module exposure concerns. For example, in cases where the candidate population was normally distributed, the bulk of examinees would likely be assigned to modules of medium difficulty, thus resulting in the need for multiple parallel modules in the middle difficulty range to alleviate high levels of

exposure. If the candidate population more generally reflected a uniform distribution, module exposure would perhaps be less of a concern for this method.

3.5.1.3 Number-Correct Scoring

The third routing strategy implemented was based on a approach using number-correct scoring, where examinee number-correct scores for the module immediately previous was the ability measure used to determine module assignment in the subsequent stage. As with the DPI method, the simplicity of implementation for development and administrative vendors may be a strong point of this methodology. However, in adapting modules to examinee ability using number-correct as a proxy for ability estimates, the process did not draw as extensively on IRT information as other methods that might be used and in effect is an under-use of available IRT information. For comparison purposes, however, this approach was of considerable interest.

To determine the number-correct cut-scores for this study, the approach used in this study involved test characteristic curves (TCCs) and consideration of the examinee ability distribution. In this strategy, for routing from Stage 1 to Stage 2, the test characteristic curve of the routing module was used to find the expected number-correct score corresponding the equal division of candidates among modules in the stage that was to be next. So, if the second stage had two modules (as in the 1-2-2 and 1-2-3), a single number-correct cut-score was needed to identify the point where approximately half of the candidates would be assigned to the easier second-stage module, and the other half would be given a harder module. A similar approach was used when the second stage

has three modules (the 1-3-3 and the 1-3-2), but in this case the number-correct score associated with dividing the simulee sample into thirds was used.

In routing from Stage 1 to Stage 2, matters were somewhat simplified by the fact that all candidates are given a module of medium difficulty in all designs evaluated here. For routing from the second to third stage, cut-scores were not only based on number-correct scores but also the difficulty level of the module taken in the second stage, which must be taken into account. In order to do this, the TCC of each module in the second stage was used to find number-correct scores that correspond to the population proportions desired. Take for example the 1-3-3 design, and consider a candidate routed to the easy module in stage two. For that candidate, routing options in the third stage consisted of an easy or medium difficulty module (to be consistent with practice, candidates would not be routed from the very easiest to the very hardest module in consecutive stages), and thus a single cut-score dividing the sample of candidates who were in the easy module in stage two in half was needed. A similar approach was used for a candidate routed to the hard Stage 2 module in the 1-3-3 design. For the candidates in the medium difficulty Stage 2 module, two cuts were required which divided that sample into thirds for assignment to the Stage 3 modules of easy, medium, and hard.

The use of TCCs and number-correct scoring was similarly used to determine cut-scores for the other design strategies under investigation in this study. Generally speaking, this method drew on both assumptions of model fit and the normal distribution of ability in the simulee population. Like other approaches, it supported the equalization of exposure of modules in all stages.

3.5.1.4 Random Module Assignment

This methodology was based on the principle of assigning candidates to modules in the second and third stages of an MST without taking the ability of the candidate or the relative difficulty of the module into account. In effect, in this method the routing of candidates was random. As implemented in this study, after simulation of the Stage 1 module, the candidate population was randomly divided among the number of modules in Stage 2. Then, again, for Stage 3, the candidate population was again randomly assigned to one of the two or three modules in that stage (depending on the design). Inclusion of this method provided a baseline for comparison of results of the other three methods implemented. This method represented a 'worst-case scenario' in which ability estimates were derived not through adaptation but a slight variation on a linear test in which sets of twenty items are selected at random for administration, although here the modules vary by difficulty and in linear testing that would not probably be the case.

3.6 Computer Simulation Method

To help clarify the procedures taken, the exact steps taken with regard to CASTISEL and MSTSIM5 in the course of the simulation for one condition (1-3-3 design, 50% increase in total information, equal partition of information across stages, DPI method, and 30% passing) are outlined below. Following each step is the section in this chapter where more detailed information about the step was provided.

1. The base target information function was specified by averaging six operational TIFs. (3.4.1)

2. For 30% passing, the base TIF was re-centered at .521 to provide maximum information at the passing score. (3.4.2)
3. To reflect a 50% increase in information, 50% more test information was added at all ability levels to the re-centered TIF. (3.4.3)
4. The TIF was divided into thirds to specify the amount of information to be obtained in each stage from each examinee. (3.4.4)
5. To specify the module difficulty differences within each stage, the TIF divided in thirds was aligned left, center, or right as needed. For Stage 1, the TIF centered at .521 was repeated three times to create three medium difficulty modules. In Stages 2 and 3, the TIF was shifted by $\frac{1}{2}$ of a standard deviation to be centered at .021 (Easy module), .521 (Medium module), and 1.021 (Hard module). (3.4.5)
6. Content balancing requirements were specified in the CASTISEL input files. (3.4.6)
7. CASTISEL was run to select items for modules within condition.
8. For MSTSIM5, the input files were specified to use the appropriate output files from CASTISEL. In the MSTSIM5 input files, denoted were the examinee and response seeds, the number of examinees in the sample (9000), the distribution of the examinee population ($\sim N(0,1)$), the number of panels (here, 3), the number of stages (3), the number of modules per stage (3), the number of items per module (20), and the cutpoints for routing¹ (with this method, -.43 and .43). (3.5)

¹ For Proximity, means of modules in stages were computed and specified. For NC, NC cut-scores were noted. For Random, values were specified and compared to random numbers during simulation.

9. MSTSIM5 was run twice for each condition to provide two replications of each condition, which allowed for decision consistency analyses to be done. The difference in the two replications was in the response seed specified in the MSTSIM5 input file.
10. The resulting simulation figures were analyzed for decision accuracy, decision consistency, ability estimation, and routing path frequency as specified in the next section.

3.7 Data Analysis

The results from the MST simulations were then analyzed with respect to several criteria of interest. The first and second outcomes of interest in this study were the levels of decision accuracy and consistency observed with these conditions at different pass rates are results with particular relevance for certification and licensure agencies interested in the measurement properties of the MST design. Thirdly, the quality of ability estimation after Stage 3 for each combination of conditions was evaluated with correlations and analysis of root mean squared errors. The fourth and last outcome of interest involves an analysis of the relative frequencies of the paths examinees are routed through, especially with regard to the different strategies for routing. These outcomes of interest, and the methods by which they were quantified, are described below.

3.7.1 Decision Accuracy

Candidates in each condition were classified as masters and non-masters based on their true and estimated abilities at three pass rates: 30%, 40%, and 50%. This provided a

range of passing rates for understanding the results of the study, and these pass rates were relevant to the agency on whose data the rest of the simulation study was modeled. True and estimated abilities above 0.521, 0.223, and 0.000 were classified as true or observed masters, while true and estimated abilities below these values were considered to be true or observed non-masters. To evaluate decision accuracy, the true and estimated classifications were then cross-tabulated to provide the proportions of correct and incorrect classifications. Within the category of incorrect classification, the proportions of Type I and II errors were considered for patterns and trends as well. In addition, true and estimated abilities after Stages 1, 2, and 3 were also correlated and reported as an indicator of the quality of ability estimation.

3.7.2 Decision Consistency

In this study, decision consistency was computed by completing two replications of each combination of conditions and comparing the classification decisions obtained for examinees across the two replications for each combination of conditions. This allowed for information about the stability of these designs and combinations of conditions to be evaluated.

3.7.3 Accuracy of Ability Estimation

Accuracy in terms of ability estimation involves the amount of error in the ability estimates observed. Such error was quantified by comparing the true abilities (which are known in this simulation study) to those observed estimates obtained by simulating the administration of a multi-stage test. As employed by Patsula (1999), the first measure of

accuracy implemented in this study was root mean squared error (RMSE). For each combination of conditions simulated in this study, the $RMSE_a$ of ability estimates was calculated for each examinee located at each of several ability levels. As shown in Equation 3.2, $RMSE_a$ was computed as:

$$RMSE_a = \sqrt{\frac{\sum_{j=1}^{n_a} (\hat{\theta}_j - \theta_j)^2}{n_a}}, \quad (3.2)$$

where $\hat{\theta}_j$ is the observed ability estimate for examinee j , θ_j is examinee j 's true ability, and number of candidates at ability level a . In this study, the nine ability levels referenced by a were intervals from -2.0 to 2.0 in increments of 0.5.

The second analysis done in the process of evaluating the quality of ability estimation in MST was a Pearson correlation between true and observed ability after the final stage of simulation for each combination of conditions. This methodology was intended to establish some notion of the strength of the relationship between true and estimated abilities for candidates in light of the different combinations of conditions implemented in this study.

3.7.4 Simulee Routing Analysis

To provide greater insight to practitioners about the nature of the routing decisions made in the process of estimating ability and making pass-fail decisions, an analysis of the proportion of candidates being routed through each possible path in each design structure was completed. The proportion of examinees taking each possible path for each of the four test structures (given each possible combination of the other variables) was designed to help provide agencies interested in using MST some

information about the specifics of the routing process and how examinees of different abilities might be routed. This analysis took on particular importance with the variable of different routing strategies, especially in exploring the ideas of module exposure and improving ability estimation.

In the 1-2-2 design, for example, recall from Figure 3.4 that there were just four possible paths a simulee can take through the three stages. Candidates could have gone Medium-Easier-Easier, Medium-Easier-Harder, Medium-Harder-Easier, or Medium-Harder-Harder (with the understanding that labels such as Easier and Harder were relative). Similarly, in the 1-3-3 design (as given in Figure 3.5 earlier), there were nine possible paths: Medium-Easier-Easier, Medium-Easier-Medium, Medium-Easier-Harder, Medium-Medium-Easier, Medium-Medium-Medium, Medium-Medium-Harder, Medium-Harder-Easier, Medium-Harder-Medium, or Medium-Harder-Harder. Of course, some of these paths were much likelier to be observed than others, as it was not likely that many examinees routed to the Easier module for Stage 2 were sent to the Harder module for Stage 3 (or vice versa), but this analysis was intended to inform about the proportions of candidates whose performance might fit that pattern in practice.

Table 3.1. Guidelines for Approximate Content Balancing Within and Across Stages

	Dimension 1	Dimension 2	Dimension 3
Original P&P forms: 60 items total	13%	37%	50%
	Approximate Proportional Counts of Items for Simulations		
Entire 60 item simulated test	8	22	30
1 st stage: 20 items	3	7	10
2 nd stage: 20 items	2	8	10
3 rd stage: 20 items	3	7	10

Figure 3.1. Test Structures of Interest

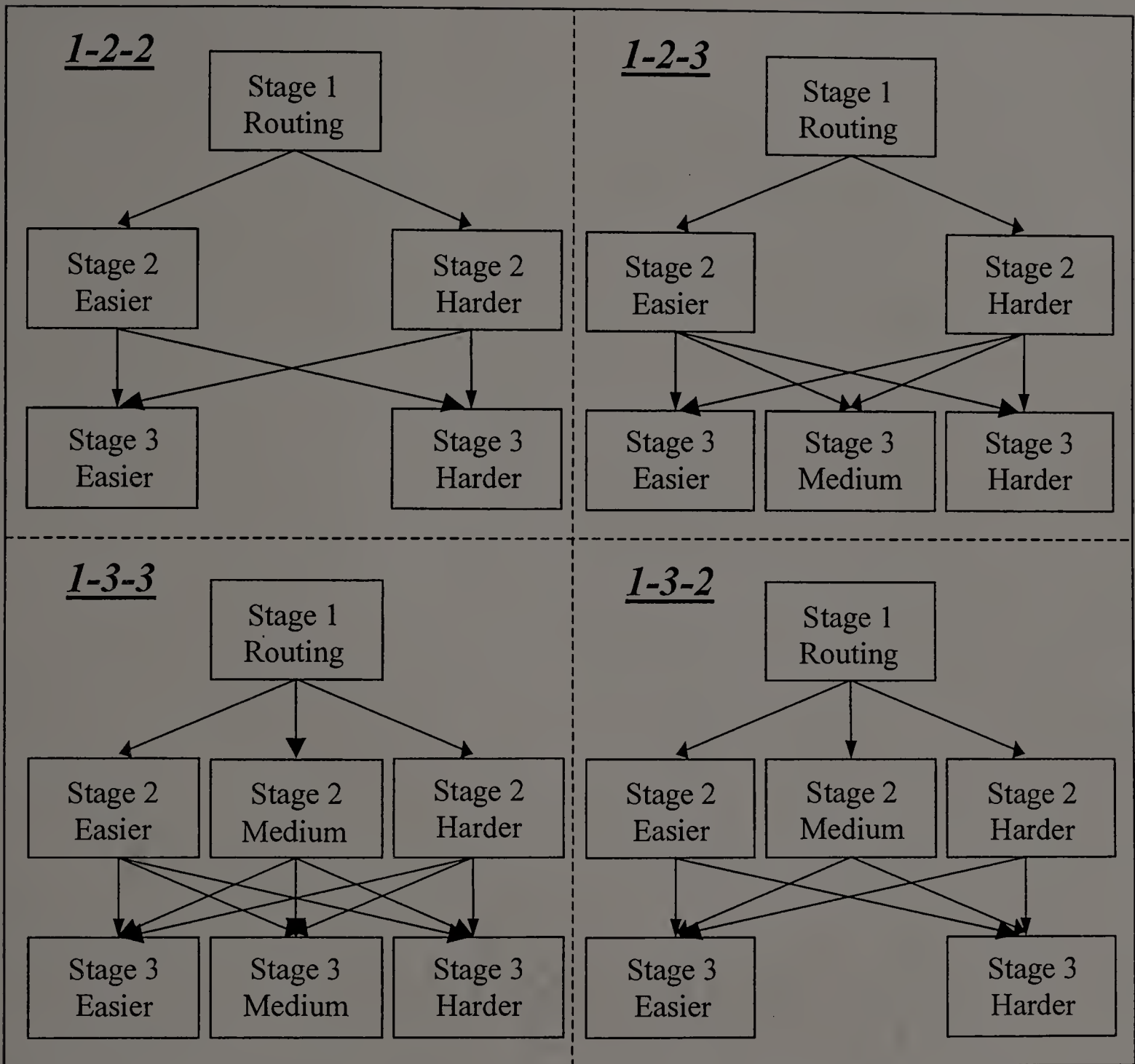


Figure 3.2. TIFs for Six Operational Forms and the Average TIF

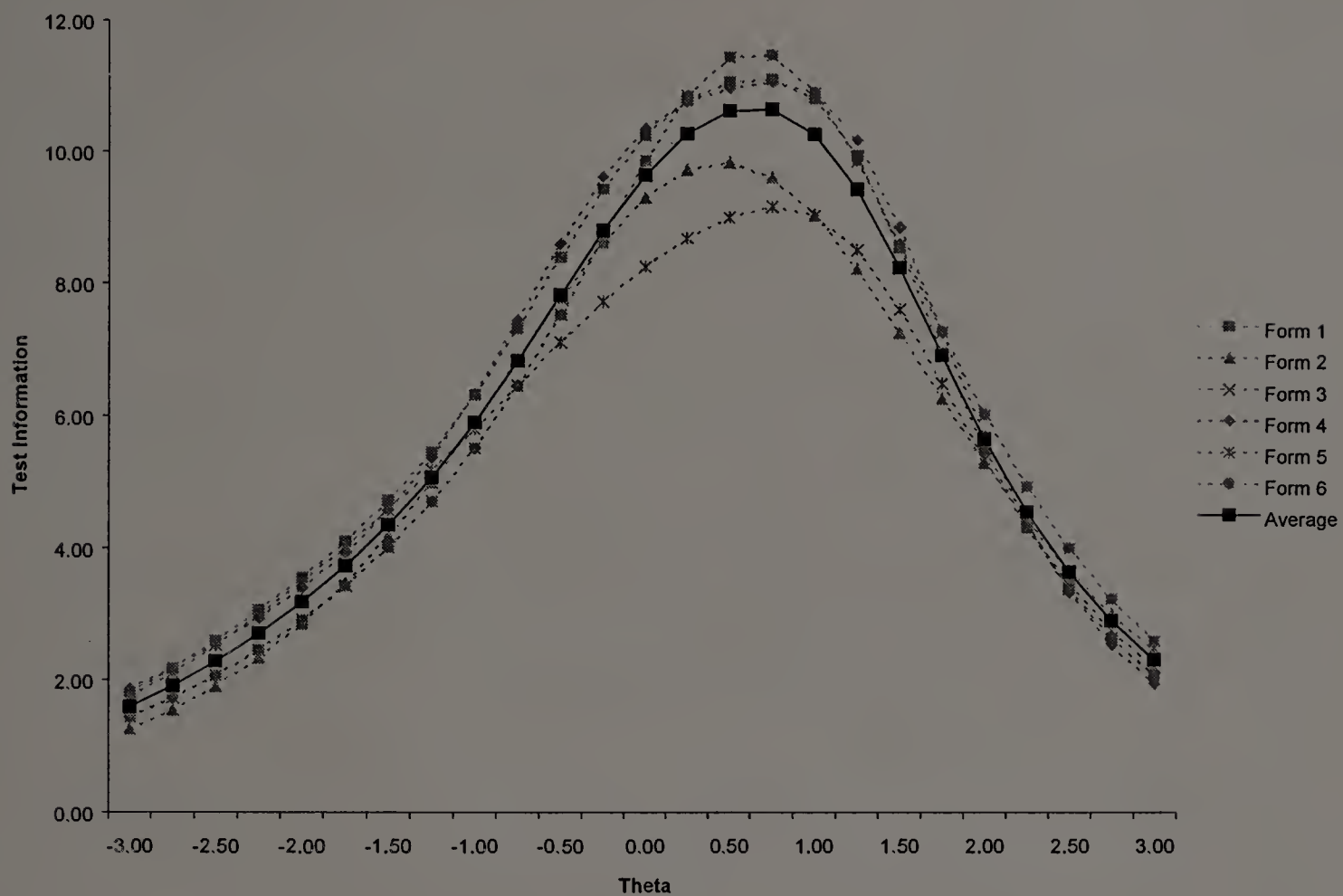


Figure 3.3. Original Average TIF and TIFs Re-centered for Three Passing Rates

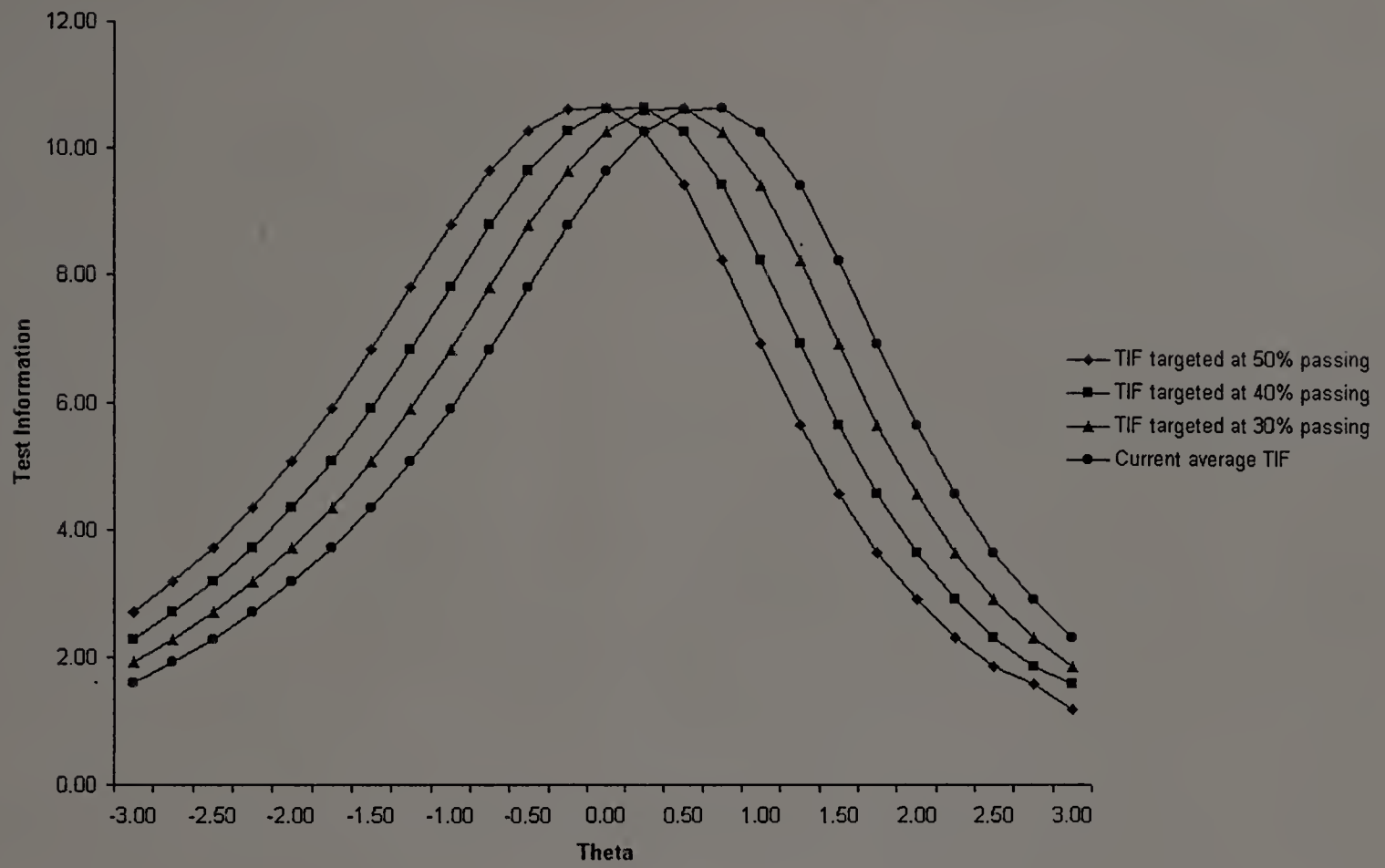


Figure 3.4. Target Test Information Functions for Three Pass Rates

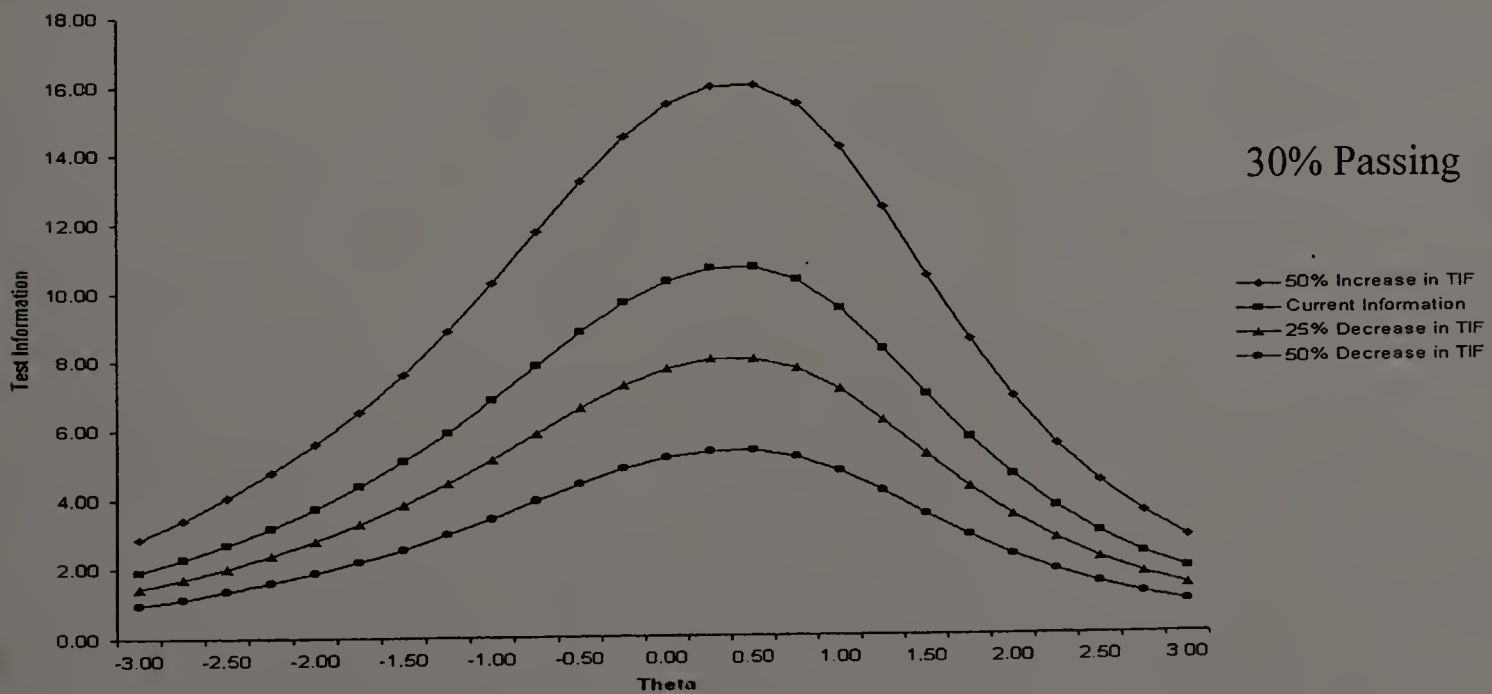
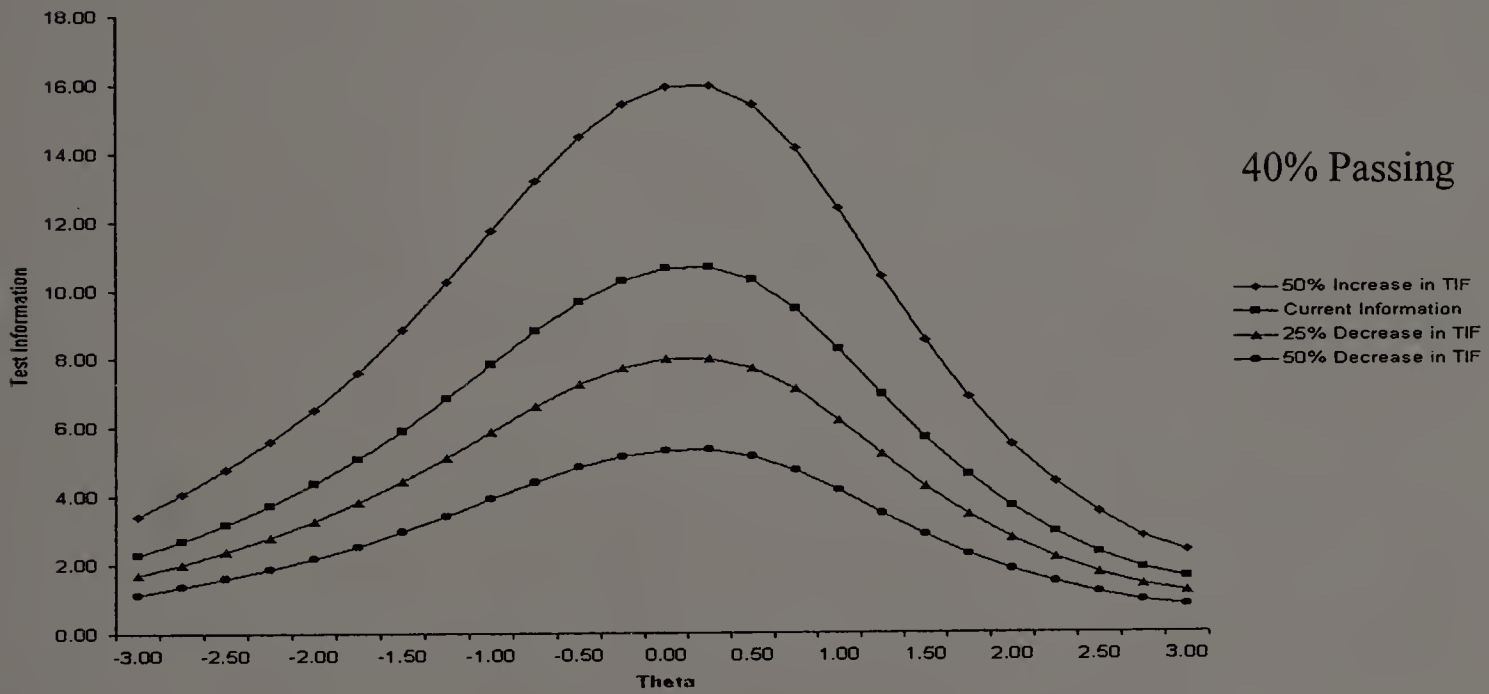
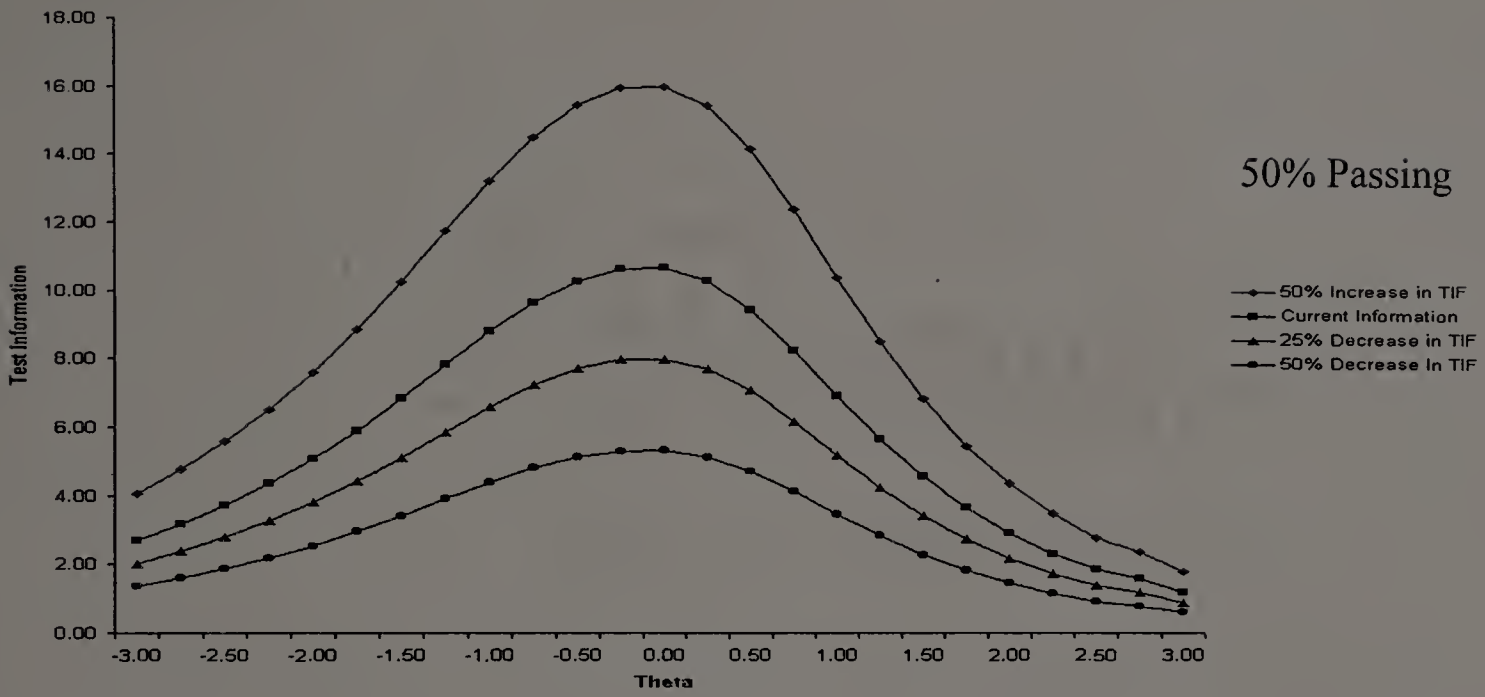
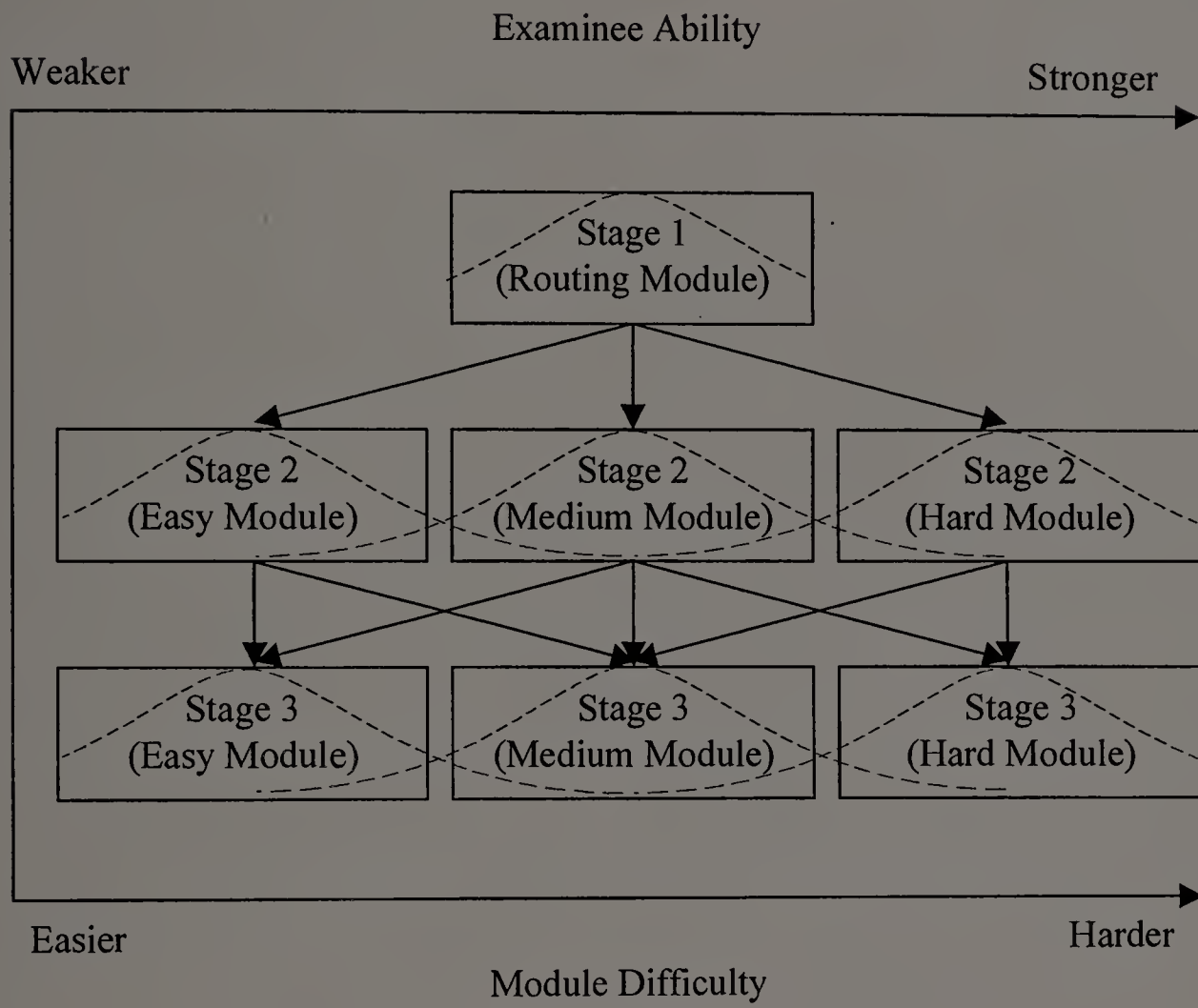


Figure 3.5. Sample Assignment of Stage-Level Information Functions to Modules



CHAPTER 4

RESULTS

4.1 Overview

In this chapter is provided an overview of the results of the simulation study described in Chapter 3. These results are presented in four parts: 1) decision accuracy, 2) decision consistency, 3) accuracy of ability estimation, and 4) routing path analysis.

4.2 Decision Accuracy

Decision accuracy, as detailed in Chapter 3, is the extent to which the decisions made using candidates' estimated abilities are consistent with decisions made based on true abilities (which are known in a simulation study). Evaluating DA with respect to the various conditions included in this study provides a sense of how well the designs (in simulation) can provide the same classification results as truth.

Tables 4.1 through 4.4 provide decision accuracy results for each of the four routing methods when the passing rate is set at 30%, these results for 40% passing are found in Tables 4.5 through 4.8, and Tables 4.9 through 4.12 present these DA results for 50% passing. In each table, DA results are included for each of the two replications completed for each condition. For each condition and replication, Tables 4.1 - 4.12 provide the percent of classification agreement, the percent of false positives and negatives observed, and the Kappa coefficient of agreement.

Overall DA results for this study were consistent with expectations in several respects. As the total amount of test information decreased (from a 50% increase to full information to 25% and 50% decreases), DA likewise decreased. Since decisions are

based on estimated abilities, less test information produces less precise ability estimates, while higher levels of test information mean that more highly informative items will be selected during module and test assembly to ensure that the higher target information functions are met, which translates into better estimation for individuals.

However, the level of this decrease does not appear constant from level to level; rather, something of an important relationship can be detected. In most conditions, the drop in DA from a 50% increase in information to full information is on the magnitude of 1.5 percentage points, meaning that approximately one and a half percent fewer decisions are consistent (between truth and the simulation) at the full information level than at the 50% increase in information level. From full information to a 25% decrease in total test information, this decline in DA is generally 1.0 to 1.5 percentage points, but from a 25% to a 50% decrease in total test information, this drop is even larger, about 2 percentage points in many conditions. When these latter two decreases are combined, the drop in decision accuracy from full information to 50% less information is larger than the drop from a 50% increase in information to full information: the differences are about 3.0 percentage points versus 1.5 percentage points. In specifying test information functions, the lesser amounts of information result in lowered DA levels, and as TIF levels decrease the decline in DA grows, but even at the 50% decrease level DA rates of 87% to 88% are observed. These results with respect to DA thus clearly provide test developers with important information about how much loss in DA could be expected with proportional reductions in test information functions at the overall test level.

As the passing rate increased from 30% to 40% and 50%, the accuracy of the decisions made declined. The most marked decline in most cases was observed for the

change from 30% to 40% passing, where the magnitude of this difference for the most part ranged from 0.5 to 1% (in some cases this difference was even larger). The difference in DA from 40% to 50% passing was more modest, about 0.4 or 0.5%. No differences in kappa, in terms of evaluating observed and expected classification, across pass rates were evident. This result reflects the nature of the interaction between the distribution of candidates and the placement of cut-scores for making pass-fail decisions, with lower DA present when the cut-score is set at a point on the ability scale where most of the examinees are.

Among the routing strategies, the DA results observed were revealing. By a slight margin, the Proximity and NC methods were associated with the highest DA levels, followed by the DPI method, which provided slightly lower DA levels than either of those routing strategies. Overall, decision accuracy was lowest for the case of Random routing, in that the random assignment of candidates to modules resulted in the lowest levels of agreement in decision classification between observed and true classifications. In contrast, since the Proximity and NC methods base routing decisions on simulee performance, they more economically use the statistical information in the adaptive test to advance the examinee through the stages of the test in the most difficulty-appropriate way. Thus, these results are significant in that it is expected that higher levels of DA would be observed when either of those approaches are implemented as compared to random or strictly population-based methods.

However, the magnitude of the DA differences between the Random method and the other three strategies was generally about one-half of one percentage point. To give that meaning, in an operational testing setting with annual testing population of 10,000

examinees, such a difference translates into perhaps 50 to 60 more misclassifications than would be seen with other routing strategies. While test developers strive for maximal accuracy in decisions, in light of DA results which across routing strategies are in the 87% to 93% range (depending on the other variables of interest), such small practical differences are indicative of generally high levels of DA that could be expected with multi-stage tests as constructed in this simulation, to the extent that many of the considerations and constraints that would be expected in credentialing and licensure practice have been integrated into this study.

With respect to the amount of test information (either an equal split across stages or a 1/2-1/4-1/4 division), clear trends to the DA results were present. At high levels of test information (either a 50% increase or full information), the equal split of information outperformed the approach where more information was gained in Stage 1. However, with less information (either a 25% or 50% decline in the size of the TIF), the 1/2-1/4-1/4 strategy was more in line with the results from the equal information method. It may be that when a greater amount of overall measurement information can be gathered, spreading such information out over stages results in more precise measurement and hence better decision-making. However, with less information overall, the results here suggest that the importance of the first routing may in some cases take on added significance in terms of the decision to be made based on observed test scores. This is to say that when lower levels of test information are specified, gathering most of that information early in the testing process from examinees may substantively improve the quality of the decisions being made. These differences were likewise reflected in the Kappa levels, whereas Kappa for higher information level conditions was higher with

equal information across stages and lower in the 1/2-1/4-1/4 conditions with less total test information.

The last variable of concern with respect to the DA results involves the choice of design structure. Across and within conditions in this study, no clear differences in the accuracy of classification differences based on design structures could be identified.

Generally speaking, the DA results on average were high across conditions. This finding is not unexpected and is likely inflated to some degree, as the model and the data are consistent with one another. In practice, programs would probably obtain somewhat lower results due to less predictable examinees, and the model-item fit would not be so precise.

4.3 Decision Consistency

The decision consistency results found in Tables 4.13, 4.14, 4.15, and 4.16 reflect DC results for each of the four routing methods. In general, the decision consistency results were found to be consistent with the findings for DA, although some interesting patterns emerged.

The DC results in this study overall ranged from approximately 90% agreement in decision classification across the two replications in each condition when high levels of test information were specified to about 80% agreement when total test information was cut by half. The better DC results were generally associated with a 30% cut score while the results obtained with a 40% or 50% cut-score were slightly lower and interestingly, the 40% and 50% cut-score results were highly consistent with one another. For the most

part, the difference in DC from the 30% cut to 40% or 50% passing was in the area of 2 to 3 percentage points.

As seen with the decision accuracy results, elevated DC results for higher levels of test information were observed with an equal distribution of test information across stages, while with lesser information the 1/2-1/4-1/4 strategy performed equally well or better. This particular pattern, however, was not evident with the random routing strategy, although it was found with DPI, Proximity, and the number-correct methods.

To continue with the results by routing strategy, DC for methods used in this simulation provided some interesting findings. Slight declines in DC were present for results from the Random strategy as compared to the other three methods, although the differences were generally small (about 1% difference). At the level of design structures, the results between designs were again largely consistent regardless of 1-2-2, 1-3-3, 1-2-3, or 1-3-2.

4.4 Accuracy of Ability Estimation

The accuracy of individual ability estimates from a test is always a major concern, even in the context of credentialing and licensure assessment where the decision outcome for each individual is the paramount outcome of interest. In this study, accuracy results for each of the 384 conditions in this study are reported with respect to 1) correlations between true abilities and final estimates of ability for candidates and 2) root mean square errors of true and estimated abilities overall and at intervals centered on nine ability levels ranging from -2.0 to 2.0.

4.4.1 Correlations between True and Estimated Abilities

In Tables 4.17, 4.18, and 4.19 are the results for these correlations at 30%, 40% and 50% passing, respectively. The correlations reported here are included as simple indicators of the strength of the relationship between the true and final estimates of ability in each condition in the simulation study. Overall, the correlations observed were quite high, from about 0.96 to a low of 0.91 or 0.90, suggesting that even in cases where less information or less optimal routing is used, the final ability estimates are reasonable approximations of candidates' true abilities.

Across conditions, several informative patterns relating to the ability estimation process with these design variables are evident. First, with respect to the implementing equivalent information across stages or a strategy with 1/2 information in Stage 1 and 1/4 information in the two subsequent stages, the results indicate that in most conditions small differences on the magnitude of approximately 0.01 to 0.03 are present depending on which division of test information was used, regardless of the other variables. This trend indicates that very slightly higher correlations between true and estimated abilities are generally associated with the practice of dividing the test information function equally among the number of stages in the test.

A second trend of note concerns the differences relating to choice of routing strategy. The random method of assigning candidates to modules in the second and third stages generally provided the lowest correlations, although the differences in the magnitude of the correlations for this routing strategy and the others were for the most part equal to 0.05. A possible explanation for this is that candidates' abilities were well estimated because of the length of the test alone (60 items), and therefore random

assignment to modules had only minimal impact on the ability estimates computed. Among the other routing strategies, the results were largely consistent, although, interestingly, the DPI method seemed to provide very slightly stronger correlations, especially in the conditions where total test information was decreased by either 25% or 50%.

Additionally, some clear differences related to the amount of test information are likewise apparent. From a 50% increase in information to full information to a 25% decrease in information, the drop in the size of the correlations at all pass rates was about 1.0 percentage points. However, with a 50% decrease in information, the increment of decrease in the correlation was more sizeable, generally about 2.5%, to about 0.9 in all conditions. This translate into a total drop in the size of the correlations from full to a 50% decrease in information of about 3 percentage points on average, which represents a considerable loss in the strength of the relationship, as compared to the negative change of about 1.0 to 1.5 % percent for the interval between a 50% increase in information and full information. The correlations were also slightly reduced for the case of 50% passing relative to 30% and 40% pass rates that were largely consistent with one another. No patterns relating the choice of MST design structure were found.

4.4.2 Root Mean Square Errors

As a second measure evaluating the accuracy of ability estimates, RMSE was assessed over all candidates in each condition (Tables 4.20, 4.21, and 4.22) and for nine ability intervals using candidates' true abilities in each condition (Figures 4.1 through

4.16). In Figures 4.1 through 4.16, each figure includes the results for both replications of each condition.

The overall RMSE results for the simulations indicated trends that were largely consistent with the correlation results. The 1/2-1/4-1/4 split of test information resulted in RMSEs that were generally slightly but noticeably higher, meaning that these ability estimates contained slightly more error than the corresponding estimates from the conditions with equivalent information across stages. RMSEs for ability estimates in the case of the highest pass rate (50%) were also slightly elevated as compared to the RMSEs observed for 30% or 40% passing rates.

These results also suggested that in many cases the DPI method of routing candidates from stage to stage performed as well or marginally better than either the proximity or the NC methods (all methods were in all conditions superior to random routing, although the magnitude of those differences were generally 0.02 to 0.03). This is an interesting result in that the DPI method in effect rank-orders candidates and assigns modules on that basis (a very norm-referenced approach), while the other two non-random routing strategies are more criterion-referenced in the way that candidates are routed, in that simulee abilities are compared to more objective standards such as the mean difficulty of the modules in the upcoming stage or number-correct cut-scores that are empirically determined by considering the statistical characteristics of the module and the ability estimate of the individual.

No consistent patterns relating to design structures were present. However, as with correlations, a trend relating differences in the magnitude of RMSE corresponding to the total amount of test information was identified. The increments from 50% increase

to full to a 25% decrease in information were consistent at all pass rates at 0.05, but from a 25% decrease to a 50% decrease in information, the size of the RMSE difference grew to 1.0.

When the RMSE results were considered at for each condition stratified into nine ability levels (as shown by the line plots in Figures 4.1 through 4.16), the results there generally supported the findings from the overall RMSEs. Consistent with centering test information at the cut-score, for all conditions RMSE was lowest in the vicinity of the cut in each condition and higher at either end of the ability scale. However, in conditions with decreasing amounts of test information and 30% passing rates, the RMSE levels exhibited a slight tendency to 'flatten out'. This is to say that at higher levels of information, RMSEs across the ability scale generally are relatively high on the tails and low in the middle region of the scale, but in many cases, when the total test information is a 50% decrease from full information and a 30% pass rate is implemented, the curves are generally much less pronounced. This has implications for practice in that if the test information function were reduced by some amount, while an overall decrease in the precision of ability estimation would be observed, such a decrease is more generally spread over the entire distribution of ability.

As expected, the higher amounts of test information were associated with lower RMSEs, and generally, equal information resulted in slightly lower RMSEs than splitting information to obtain half in Stage 1. Across routing methods, no RMSE differences could be detected.

4.5 Routing Path Analysis

In reviewing the frequencies of the paths taken across conditions, it was apparent that no differences were present related to the choice of pass rate. For this reason, the results presented in Tables 4.23 through 4.26 are averages of the percentages of candidates being routed in each path across pass rates and the two replications.

These results were illuminating in several respects. Each of the four tables provides an average percent of candidates taking each of the possible paths for one of the four design structures under consideration in this study, and these results are broken out within each table by routing strategy. In the 1-2-2 design structure, there were four possible paths, and seven in the 1-3-3. Both the 1-2-3 and the 1-3-2 had six possible paths to which candidates could have been assigned.

4.5.1 Routing Path Analysis for the 1-2-2 Design Structure

In the context of the 1-2-2 design structure (Table 4.23), the Random method (as intended) assigned candidates to paths in equal proportions, and the DPI method resulted in relatively low proportions of candidates being assigned to modules of different difficulty levels for the second and third stages. As information decreased, for all routing strategies except Random the number of candidates whose module difficulty levels changed between stages increased. In practical terms, as module information is lessened, more error is present in the ability estimation process, and for candidates in the vicinity of the cut-scores for routing, the likelihood of their being routed to one module or another increased because their estimated ability is less precise (i.e., more inconsistent with the true ability).

In terms of the division of the test information function, the results here too varied in an interesting way, although the implications were quite different across routing strategies. With the DPI method, using equal information across stages resulted in slightly more examinees changing module difficulty between the second and third stages regardless of overall amount of test information, as compared to the approach where half of the test information is specified in the first stage and a quarter in each of the later stages.

4.5.2 Routing Path Analysis for the 1-3-3 Design Structure

As shown in Table 4.24, the general trend to the frequencies of candidates taking each path for this design indicated that for the DPI, Routing, and NC methods, a large proportion of candidates were routed to modules of equivalent difficulty in the second and third stages. This is to say that if they were routed to an easy module for Stage 2, then they similarly received an easy module in Stage 3 (and so on for medium and hard modules and paths). Indeed, depending on the condition, approximately 70% to 80% of candidates were routed in this way. Though a large proportion of examinees were routed in this way, 20% to 30% of examinees did receive modules of different difficulty from Stage 2 to Stage 3.

In these results, no patterns relating to the conditions in the simulation could be detected. However, with respect to exposure of modules in each stage of the MST, regardless of routing strategy used, exposure levels for individual modules were largely consistent, which is good news for practice.

4.5.3 Routing Path Analysis for the 1-2-3 Design Structure

Important to notice in this design, illustrated by the results in Table 4.25, is that relatively few examinees were routed from the easier Stage 2 module to the hardest Stage 3 module, or, conversely, the harder Stage 2 module to the easiest Stage 3 module. In operational testing, such large 'jumps' in ability are often flagged as aberrant and may be indicative of some problem with ability estimation (either inappropriate behavior on the part of the examinee or a technical difficulty with the test itself). Across conditions, perhaps 60% of candidates were administered modules of equivalent difficulty in the second and third stages.

4.5.4 Routing Path Analysis for the 1-3-2 Design Structure

For this design, as with the 1-2-3, it is evident that proportionally many more candidates were routed to easy-easy or hard-hard in Stages 2 and 3. Again, fewer examinees seemed to have been routed to modules of different difficulties between the later two stages of the MST, but still on average 30% were.

4.6 Summary

Across analyses reported here, the results were largely consistent in their implications for operational multi-stage testing. The most unexpected results obtained in this simulation concerns the choice of how to divide test information among stages: across conditions and analyses, the results indicate that with high overall amounts of information the preferable approach is to spilt information equally. When lesser levels of information are to be used, better results both with respect to basic ability estimation and

making pass-fail decisions using those estimates are likely to be obtained through a strategy in which more of that information is gathered in the earlier part of the multi-stage test (translating into more efficient routing of candidates through the MST structure in the absence of higher levels of information).

In addition, while clear differences in the results were observed between the levels of information, the most sizeable differences concerning decisions and ability estimation were noted in moving from full information or a 25% decrease in information to a 50% decrease in information. Such declines in DA, DC, and the accuracy of ability estimation between levels of test information have clear repercussions for individuals with respect to the quality of the measurement results in this high-stakes context. To the extent that test developers are able to specify and meet high levels of test information, the measurement outcomes of interest are likely to be psychometrically sound. However, when item development or other operational considerations constrain or negatively impact test assembly, then understanding the trade-off in measurement precision that can be expected becomes necessary. In this case, two percentage points' worth of DA might be lost when information is decreased by half: if a program tests 100,000 candidates for certification per year, that translates into 2,000 more misclassifications. If the decrease in DC is 5 percentage points, that is 5,000 examinees whose decision classifications from one test occasion to another would vary one way or another. In high-stakes testing, it is these people for whom maximizing measurement precision is critical.

Interestingly, no substantive differences between the routing strategies adopted were evident. Aided by the finding that even the Random strategy provided comparable results in all respects, the length of the test may be such that ability estimation is already

approaching such a high level of precision that with so many items per stage (and thus across the entire test) the exact approach taken may matter less than simply administering many items to the individual, within the general MST structure and method.

As no differences due to routing strategy were patent, so too was there an absence of differences in the results due to design structure. The choice of two or three modules in the second and third stages seemed to have no significant impact on the results in any respect.

Table 4.1. Decision Accuracy for DPI Routing at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	92.8	3.8	3.4	0.829	92.7	4.1	3.2	0.829
	1-2-2	Full	92.0	3.7	4.3	0.829	90.6	4.7	4.7	0.805
	1-2-2	25% dec.	90.0	4.7	5.3	0.766	90.2	5.4	4.4	0.770
	1-2-2	50% dec.	87.8	6.8	5.5	0.714	88.0	6.7	5.3	0.719
	1-3-3	50% inc.	92.9	3.8	3.3	0.833	92.7	4.0	3.3	0.828
	1-3-3	Full	91.9	4.8	3.3	0.833	91.2	4.8	4.0	0.793
	1-3-3	25% dec.	90.4	5.2	4.4	0.775	90.1	5.4	4.5	0.768
	1-3-3	50% dec.	87.6	6.9	5.5	0.710	88.3	6.6	5.1	0.726
	1-2-3	50% inc.	93.0	3.8	3.2	0.835	92.9	3.8	3.3	0.834
	1-2-3	Full	91.7	3.8	4.2	0.835	91.7	4.7	3.7	0.804
	1-2-3	25% dec.	89.8	5.7	4.4	0.762	90.0	5.6	4.4	0.766
	1-2-3	50% dec.	88.4	6.4	5.2	0.727	87.9	7.0	5.1	0.719
	1-3-2	50% inc.	93.0	3.8	3.2	0.835	92.5	4.3	3.2	0.824
	1-3-2	Full	92.2	4.0	3.8	0.835	91.5	4.6	3.9	0.800
	1-3-2	25% dec.	90.2	5.3	4.5	0.770	90.2	5.5	4.3	0.771
	1-3-2	50% dec.	87.6	7.0	5.4	0.711	87.7	7.0	5.2	0.714
2	1-2-2	50% inc.	92.7	4.3	3.1	0.828	92.6	4.0	3.3	0.827
	1-2-2	Full	91.0	5.1	3.9	0.790	89.8	5.4	4.7	0.791
	1-2-2	25% dec.	89.9	4.4	5.7	0.763	90.0	5.8	4.2	0.766
	1-2-2	50% dec.	88.1	6.8	5.1	0.722	87.6	7.2	5.1	0.713
	1-3-3	50% inc.	92.6	4.0	3.4	0.825	92.9	3.6	3.4	0.833
	1-3-3	Full	90.9	5.0	4.1	0.785	91.2	4.9	3.9	0.793
	1-3-3	25% dec.	90.2	5.6	4.2	0.770	89.8	5.7	4.5	0.762
	1-3-3	50% dec.	87.3	7.3	5.4	0.704	87.9	6.9	5.3	0.717
	1-2-3	50% inc.	92.9	3.9	3.2	0.832	92.5	4.1	3.4	0.823
	1-2-3	Full	91.0	4.9	4.1	0.788	91.0	5.1	3.9	0.790
	1-2-3	25% dec.	90.0	5.5	4.4	0.766	90.2	5.5	4.3	0.771
	1-2-3	50% dec.	87.4	7.2	5.4	0.706	87.8	6.9	5.2	0.716
	1-3-2	50% inc.	93.0	3.8	3.2	0.836	92.9	3.8	3.3	0.832
	1-3-2	Full	91.3	4.9	3.8	0.796	91.2	4.9	3.9	0.793
	1-3-2	25% dec.	89.9	5.8	4.4	0.763	89.8	5.8	4.4	0.762
	1-3-2	50% dec.	87.4	7.2	5.4	0.706	87.6	7.3	5.1	0.713

Table 4.2. Decision Accuracy for Proximity Routing at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	93.3	3.5	3.2	0.842	93.0	3.9	3.2	0.835
	1-2-2	Full	91.9	4.1	4.0	0.808	90.1	5.5	4.5	0.796
	1-2-2	25% dec.	90.3	4.5	5.2	0.773	90.3	5.4	4.2	0.774
	1-2-2	50% dec.	88.0	6.6	5.4	0.719	88.1	6.6	5.3	0.722
	1-3-3	50% inc.	93.2	3.6	3.2	0.840	92.8	3.9	3.3	0.829
	1-3-3	Full	91.6	4.5	4.0	0.801	91.8	4.5	3.7	0.807
	1-3-3	25% dec.	90.5	5.1	4.4	0.776	90.2	5.5	4.4	0.769
	1-3-3	50% dec.	88.2	6.6	5.2	0.725	88.2	6.7	5.1	0.724
	1-2-3	50% inc.	92.9	3.7	3.4	0.833	92.8	3.8	3.4	0.831
	1-2-3	Full	91.4	4.6	4.0	0.798	91.3	4.8	3.9	0.796
	1-2-3	25% dec.	90.2	5.3	4.5	0.770	90.0	5.4	4.6	0.766
	1-2-3	50% dec.	88.1	6.9	5.0	0.722	88.3	6.5	5.2	0.727
	1-3-2	50% inc.	93.1	3.6	3.3	0.838	92.6	3.9	3.6	0.825
	1-3-2	Full	91.6	4.5	3.9	0.803	91.6	4.6	3.8	0.803
	1-3-2	25% dec.	90.4	5.2	4.5	0.773	90.2	5.6	4.2	0.771
	1-3-2	50% dec.	88.3	6.5	5.1	0.727	88.1	6.8	5.0	0.724
2	1-2-2	50% inc.	92.9	4.0	3.1	0.834	92.6	4.1	3.4	0.825
	1-2-2	Full	91.3	4.8	3.9	0.797	90.4	5.1	4.5	0.803
	1-2-2	25% dec.	90.3	4.2	5.5	0.773	89.7	5.8	4.5	0.758
	1-2-2	50% dec.	88.0	6.8	5.2	0.719	88.2	6.9	4.9	0.725
	1-3-3	50% inc.	92.7	4.0	3.3	0.829	92.5	4.3	3.2	0.825
	1-3-3	Full	91.7	4.6	3.7	0.804	91.4	4.7	3.9	0.798
	1-3-3	25% dec.	90.1	5.5	4.3	0.769	89.9	5.6	4.5	0.762
	1-3-3	50% dec.	87.7	7.0	5.3	0.714	88.3	6.6	5.1	0.727
	1-2-3	50% inc.	92.9	4.0	3.1	0.833	92.5	4.2	3.3	0.823
	1-2-3	Full	91.5	4.7	3.9	0.799	91.5	4.7	3.8	0.801
	1-2-3	25% dec.	90.0	5.7	4.4	0.765	89.7	5.9	4.4	0.759
	1-2-3	50% dec.	87.7	7.0	5.2	0.715	88.1	6.8	5.1	0.723
	1-3-2	50% inc.	92.9	4.0	3.1	0.833	92.5	4.3	3.3	0.824
	1-3-2	Full	91.3	4.9	3.8	0.796	91.4	4.7	3.9	0.798
	1-3-2	25% dec.	90.0	5.6	4.4	0.766	89.9	5.7	4.4	0.765
	1-3-2	50% dec.	87.8	7.0	5.2	0.717	88.1	6.7	5.2	0.722

Table 4.3. Decision Accuracy for Number-Correct Routing at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	93.3	3.5	3.2	0.841	93.0	3.9	3.1	0.835
	1-2-2	Full	91.9	4.2	4.0	0.808	90.1	5.5	4.5	0.796
	1-2-2	25% dec.	90.3	4.4	5.3	0.772	90.3	5.4	4.3	0.772
	1-2-2	50% dec.	87.9	6.7	5.4	0.718	88.2	6.5	5.3	0.725
	1-3-3	50% inc.	93.1	3.7	3.2	0.837	92.8	3.8	3.4	0.830
	1-3-3	Full	91.5	4.4	4.1	0.800	91.8	4.6	3.6	0.807
	1-3-3	25% dec.	90.5	5.1	4.4	0.776	90.1	5.4	4.5	0.767
	1-3-3	50% dec.	88.1	6.6	5.2	0.722	88.2	6.6	5.2	0.725
	1-2-3	50% inc.	93.1	3.6	3.3	0.837	92.7	3.9	3.4	0.829
	1-2-3	Full	91.6	4.4	4.0	0.802	91.4	4.7	3.9	0.797
	1-2-3	25% dec.	90.3	5.3	4.4	0.772	90.0	5.5	4.5	0.765
	1-2-3	50% dec.	88.2	6.7	5.1	0.725	88.5	6.4	5.1	0.730
	1-3-2	50% inc.	93.0	3.7	3.3	0.834	92.6	3.8	3.6	0.824
	1-3-2	Full	91.7	4.4	3.8	0.805	91.7	4.5	3.8	0.804
	1-3-2	25% dec.	90.4	5.1	4.5	0.773	90.2	5.4	4.4	0.770
	1-3-2	50% dec.	88.4	6.5	5.1	0.730	88.0	6.8	5.2	0.720
2	1-2-2	50% inc.	93.0	4.0	3.1	0.835	92.6	4.0	3.3	0.826
	1-2-2	Full	91.4	4.8	3.8	0.798	90.5	5.1	4.3	0.805
	1-2-2	25% dec.	90.4	4.2	5.5	0.774	89.6	5.8	4.6	0.757
	1-2-2	50% dec.	88.0	6.7	5.3	0.719	88.2	6.9	4.9	0.726
	1-3-3	50% inc.	92.5	4.0	3.4	0.824	92.5	4.3	3.2	0.823
	1-3-3	Full	91.6	4.6	3.7	0.804	91.3	4.8	3.9	0.795
	1-3-3	25% dec.	90.1	5.6	4.3	0.769	89.7	5.7	4.6	0.757
	1-3-3	50% dec.	87.7	7.0	5.4	0.712	88.4	6.7	5.0	0.729
	1-2-3	50% inc.	92.9	4.0	3.1	0.833	92.5	4.2	3.4	0.823
	1-2-3	Full	91.3	4.8	3.9	0.795	91.6	4.6	3.8	0.801
	1-2-3	25% dec.	90.0	5.7	4.3	0.766	89.8	5.7	4.5	0.761
	1-2-3	50% dec.	88.0	6.9	5.1	0.721	88.1	6.7	5.1	0.723
	1-3-2	50% inc.	92.9	3.9	3.2	0.834	92.4	4.3	3.3	0.822
	1-3-2	Full	91.4	4.9	3.8	0.798	91.3	4.7	3.9	0.796
	1-3-2	25% dec.	90.0	5.6	4.3	0.767	89.7	5.7	4.5	0.760
	1-3-2	50% dec.	87.7	7.0	5.3	0.714	88.0	6.7	5.2	0.720

Table 4.4. Decision Accuracy for Random Routing at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	92.3	4.2	3.5	0.820	92.9	4.0	3.1	0.832
	1-2-2	Full	91.5	4.8	3.8	0.800	90.1	5.5	4.4	0.796
	1-2-2	25% dec.	89.9	4.5	5.6	0.764	89.0	6.3	4.7	0.743
	1-2-2	50% dec.	86.9	7.3	5.8	0.694	87.7	7.0	5.3	0.713
	1-3-3	50% inc.	92.5	4.1	3.4	0.824	92.3	4.1	3.6	0.818
	1-3-3	Full	91.0	4.7	4.3	0.788	91.2	5.0	3.9	0.793
	1-3-3	25% dec.	90.0	5.7	4.3	0.766	89.9	5.8	4.3	0.765
	1-3-3	50% dec.	87.6	7.2	5.1	0.713	87.6	7.2	5.2	0.712
	1-2-3	50% inc.	92.5	4.1	3.5	0.823	92.4	4.3	3.3	0.821
	1-2-3	Full	90.9	5.2	4.0	0.786	91.3	5.0	3.7	0.797
	1-2-3	25% dec.	89.4	6.0	4.5	0.753	89.8	5.6	4.7	0.759
	1-2-3	50% dec.	87.2	7.4	5.4	0.702	87.9	7.1	5.0	0.719
	1-3-2	50% inc.	92.6	3.9	3.5	0.826	92.7	4.0	3.3	0.829
	1-3-2	Full	90.8	5.3	3.9	0.784	90.8	5.1	4.1	0.784
	1-3-2	25% dec.	90.0	5.6	4.4	0.766	90.0	5.8	4.2	0.766
	1-3-2	50% dec.	87.4	7.3	5.4	0.706	87.7	6.8	5.5	0.712
2	1-2-2	50% inc.	92.3	4.1	3.5	0.819	92.4	4.1	3.4	0.822
	1-2-2	Full	91.0	5.1	3.9	0.790	90.1	5.5	4.4	0.796
	1-2-2	25% dec.	89.9	4.5	5.6	0.763	89.8	5.7	4.6	0.760
	1-2-2	50% dec.	87.7	7.0	5.4	0.712	88.1	7.0	4.9	0.723
	1-3-3	50% inc.	92.6	4.0	3.4	0.827	92.3	4.1	3.6	0.819
	1-3-3	Full	91.0	5.1	3.9	0.790	91.2	4.7	4.2	0.792
	1-3-3	25% dec.	89.5	5.7	4.8	0.753	90.0	5.7	4.3	0.766
	1-3-3	50% dec.	87.5	7.2	5.3	0.710	87.2	7.6	5.2	0.702
	1-2-3	50% inc.	92.8	4.0	3.1	0.832	92.7	3.9	3.4	0.827
	1-2-3	Full	91.4	5.0	3.6	0.800	91.1	5.0	4.0	0.790
	1-2-3	25% dec.	89.6	5.8	4.6	0.756	90.0	5.9	4.2	0.766
	1-2-3	50% dec.	87.3	7.2	5.5	0.704	88.1	6.7	5.3	0.721
	1-3-2	50% inc.	92.5	3.9	3.6	0.823	92.7	3.9	3.4	0.828
	1-3-2	Full	90.8	5.3	3.9	0.785	91.3	4.7	4.0	0.795
	1-3-2	25% dec.	89.0	6.0	5.0	0.743	90.1	5.7	4.1	0.770
	1-3-2	50% dec.	87.0	7.5	5.5	0.697	88.1	6.8	5.1	0.722

Table 4.5. Decision Accuracy for DPI Routing at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	92.5	3.8	3.7	0.845	91.7	4.1	4.2	0.829
	1-2-2	Full	89.5	4.9	5.6	0.790	89.6	5.0	5.5	0.791
	1-2-2	25% dec.	88.6	5.9	5.5	0.766	89.5	5.4	5.1	0.784
	1-2-2	50% dec.	86.7	7.0	6.3	0.727	86.7	7.0	6.3	0.726
	1-3-3	50% inc.	92.6	3.9	3.5	0.848	91.8	4.3	3.9	0.832
	1-3-3	Full	90.4	5.0	4.6	0.802	90.7	4.8	4.5	0.808
	1-3-3	25% dec.	88.5	5.9	5.6	0.762	89.4	5.6	5.0	0.781
	1-3-3	50% dec.	86.7	6.9	6.4	0.727	86.6	7.2	6.1	0.725
	1-2-3	50% inc.	92.1	4.1	3.8	0.837	91.7	4.1	4.1	0.830
	1-2-3	Full	90.5	4.7	4.9	0.803	90.8	4.8	4.5	0.810
	1-2-3	25% dec.	88.9	5.8	5.3	0.771	89.2	5.8	5.0	0.778
	1-2-3	50% dec.	87.0	6.7	6.3	0.731	87.1	6.8	6.1	0.734
	1-3-2	50% inc.	92.5	3.9	3.6	0.846	92.1	4.4	3.5	0.838
	1-3-2	Full	90.2	5.1	4.7	0.798	90.6	4.8	4.5	0.807
	1-3-2	25% dec.	88.6	6.0	5.4	0.766	89.3	5.8	4.9	0.779
	1-3-2	50% dec.	86.5	7.2	6.4	0.721	86.6	7.0	6.4	0.725
2	1-2-2	50% inc.	91.9	4.4	3.7	0.833	91.6	4.6	3.8	0.828
	1-2-2	Full	89.5	5.4	5.1	0.790	89.4	5.1	5.5	0.789
	1-2-2	25% dec.	88.9	5.9	5.2	0.771	88.9	6.1	5.0	0.773
	1-2-2	50% dec.	86.8	7.1	6.1	0.728	86.8	7.1	6.1	0.728
	1-3-3	50% inc.	92.4	4.1	3.5	0.843	91.6	4.6	3.7	0.828
	1-3-3	Full	90.2	5.3	4.4	0.799	90.2	5.2	4.6	0.799
	1-3-3	25% dec.	88.7	6.0	5.3	0.767	88.8	6.0	5.2	0.770
	1-3-3	50% dec.	86.7	7.1	6.3	0.725	86.6	7.1	6.3	0.725
	1-2-3	50% inc.	91.8	4.3	3.9	0.831	91.5	4.6	3.9	0.825
	1-2-3	Full	90.4	5.2	4.4	0.802	90.3	5.3	4.4	0.801
	1-2-3	25% dec.	88.8	5.9	5.3	0.770	89.1	5.7	5.2	0.775
	1-2-3	50% dec.	86.6	7.4	6.0	0.724	86.1	7.6	6.3	0.715
	1-3-2	50% inc.	92.3	3.9	3.8	0.842	91.9	4.3	3.8	0.832
	1-3-2	Full	90.7	4.8	4.5	0.809	90.1	5.2	4.7	0.796
	1-3-2	25% dec.	88.7	6.1	5.2	0.767	88.7	6.0	5.3	0.768
	1-3-2	50% dec.	86.4	7.2	6.4	0.721	86.1	7.1	6.8	0.713

Table 4.6. Decision Accuracy for Proximity Routing at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.8	4.6	3.6	0.832	91.9	4.4	3.7	0.833
	1-2-2	Full	89.7	5.1	5.2	0.794	90.2	4.8	4.9	0.805
	1-2-2	25% dec.	88.7	6.0	5.3	0.768	88.9	5.9	5.1	0.772
	1-2-2	50% dec.	86.5	7.4	6.2	0.722	86.2	7.5	6.2	0.717
	1-3-3	50% inc.	92.4	4.2	3.4	0.843	91.6	4.5	3.9	0.826
	1-3-3	Full	90.3	5.1	4.6	0.799	90.4	5.2	4.4	0.802
	1-3-3	25% dec.	88.7	6.2	5.1	0.768	88.9	6.1	5.0	0.772
	1-3-3	50% dec.	86.7	7.3	6.0	0.727	86.6	7.2	6.3	0.724
	1-2-3	50% inc.	92.3	4.1	3.6	0.842	91.6	4.4	4.1	0.826
	1-2-3	Full	90.3	5.2	4.5	0.801	90.0	5.5	4.5	0.794
	1-2-3	25% dec.	88.7	6.0	5.3	0.768	89.0	5.8	5.2	0.773
	1-2-3	50% dec.	86.1	7.5	6.5	0.713	86.5	7.4	6.1	0.722
	1-3-2	50% inc.	92.0	4.3	3.7	0.835	91.5	4.5	4.0	0.825
	1-3-2	Full	89.9	5.3	4.7	0.793	90.2	5.3	4.5	0.799
	1-3-2	25% dec.	88.8	6.0	5.2	0.769	88.8	6.0	5.3	0.769
	1-3-2	50% dec.	86.0	7.7	6.3	0.712	86.6	7.3	6.1	0.724
2	1-2-2	50% inc.	91.9	4.4	3.7	0.833	91.6	4.3	4.1	0.827
	1-2-2	Full	89.9	5.1	5.0	0.798	89.9	5.1	5.1	0.797
	1-2-2	25% dec.	88.6	6.2	5.2	0.766	88.8	6.0	5.2	0.770
	1-2-2	50% dec.	86.6	7.4	6.0	0.725	86.2	7.4	6.4	0.716
	1-3-3	50% inc.	91.7	4.4	3.9	0.830	91.8	4.3	3.9	0.830
	1-3-3	Full	90.3	5.3	4.4	0.800	90.3	5.2	4.5	0.800
	1-3-3	25% dec.	89.1	5.9	5.0	0.775	88.7	6.1	5.2	0.768
	1-3-3	50% dec.	86.3	7.5	6.2	0.719	86.2	7.8	6.0	0.717
	1-2-3	50% inc.	91.7	4.4	3.9	0.829	91.8	4.2	4.0	0.832
	1-2-3	Full	90.2	5.2	4.6	0.799	90.2	5.2	4.6	0.797
	1-2-3	25% dec.	89.0	5.9	5.1	0.775	88.6	6.2	5.2	0.765
	1-2-3	50% dec.	86.5	7.3	6.3	0.722	86.3	7.5	6.2	0.719
	1-3-2	50% inc.	92.1	4.3	3.7	0.837	91.9	4.2	3.9	0.833
	1-3-2	Full	90.4	5.3	4.3	0.802	90.3	5.2	4.5	0.800
	1-3-2	25% dec.	89.0	6.0	5.0	0.774	88.8	6.1	5.1	0.771
	1-3-2	50% dec.	86.4	7.6	6.0	0.721	86.3	7.7	6.0	0.720

Table 4.7. Decision Accuracy for Number-Correct Routing at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.9	4.4	3.7	0.833	91.8	4.4	3.9	0.830
	1-2-2	Full	89.5	4.9	5.5	0.791	89.9	4.9	5.2	0.798
	1-2-2	25% dec.	88.7	6.0	5.3	0.768	89.1	5.7	5.2	0.776
	1-2-2	50% dec.	86.4	7.3	6.2	0.721	86.3	7.4	6.3	0.718
	1-3-3	50% inc.	92.2	4.2	3.5	0.840	91.6	4.5	3.9	0.827
	1-3-3	Full	90.1	5.1	4.8	0.796	90.4	5.2	4.4	0.802
	1-3-3	25% dec.	88.6	6.3	5.0	0.766	89.0	6.0	5.0	0.774
	1-3-3	50% dec.	86.4	7.2	6.3	0.721	86.7	7.1	6.2	0.726
	1-2-3	50% inc.	92.1	4.2	3.7	0.837	91.6	4.4	4.0	0.826
	1-2-3	Full	90.3	5.3	4.4	0.801	89.6	5.8	4.7	0.785
	1-2-3	25% dec.	89.0	5.9	5.1	0.773	88.7	6.0	5.3	0.768
	1-2-3	50% dec.	86.0	7.4	6.6	0.712	86.6	7.3	6.1	0.724
	1-3-2	50% inc.	92.0	4.3	3.7	0.836	91.4	4.5	4.1	0.823
	1-3-2	Full	89.8	5.4	4.9	0.789	90.1	5.4	4.5	0.797
	1-3-2	25% dec.	88.5	6.1	5.4	0.763	88.9	5.9	5.2	0.772
	1-3-2	50% dec.	86.0	7.6	6.4	0.712	86.6	7.3	6.1	0.725
2	1-2-2	50% inc.	91.9	4.2	3.9	0.833	91.6	4.3	4.0	0.828
	1-2-2	Full	89.9	5.1	5.0	0.798	90.0	5.1	5.0	0.799
	1-2-2	25% dec.	88.5	6.2	5.3	0.763	88.7	6.0	5.3	0.768
	1-2-2	50% dec.	86.5	7.2	6.3	0.722	86.0	7.4	6.5	0.713
	1-3-3	50% inc.	91.8	4.3	3.9	0.832	91.8	4.2	4.0	0.830
	1-3-3	Full	89.9	5.6	4.6	0.792	90.2	5.3	4.5	0.799
	1-3-3	25% dec.	88.7	6.3	5.0	0.768	88.7	6.1	5.2	0.767
	1-3-3	50% dec.	86.5	7.4	6.1	0.722	86.2	7.7	6.1	0.717
	1-2-3	50% inc.	91.8	4.2	4.0	0.831	91.9	4.1	4.0	0.832
	1-2-3	Full	90.3	5.0	4.7	0.800	90.0	5.3	4.7	0.794
	1-2-3	25% dec.	88.7	6.1	5.2	0.767	88.5	6.2	5.3	0.763
	1-2-3	50% dec.	86.1	7.7	6.2	0.715	86.0	7.7	6.3	0.712
	1-3-2	50% inc.	91.9	4.3	3.8	0.834	91.9	4.1	4.0	0.833
	1-3-2	Full	90.2	5.3	4.5	0.798	90.3	5.3	4.4	0.800
	1-3-2	25% dec.	88.7	6.2	5.1	0.769	88.9	5.9	5.2	0.771
	1-3-2	50% dec.	86.4	7.6	6.0	0.721	86.1	7.8	6.2	0.714

Table 4.8. Decision Accuracy for Random Routing at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	92.0	4.3	3.6	0.836	91.6	4.5	3.9	0.826
	1-2-2	Full	89.3	5.4	5.2	0.786	89.8	5.3	4.9	0.796
	1-2-2	25% dec.	89.0	5.9	5.1	0.773	88.7	5.9	5.4	0.767
	1-2-2	50% dec.	85.7	7.8	6.5	0.707	86.1	7.3	6.6	0.713
	1-3-3	50% inc.	91.7	4.6	3.7	0.830	91.5	4.7	3.8	0.825
	1-3-3	Full	90.3	5.3	4.4	0.801	90.7	5.0	4.3	0.809
	1-3-3	25% dec.	88.4	6.5	5.1	0.761	89.1	5.8	5.1	0.776
	1-3-3	50% dec.	86.1	7.5	6.4	0.715	86.6	7.5	5.9	0.724
	1-2-3	50% inc.	91.8	4.4	3.8	0.832	91.8	4.5	3.8	0.831
	1-2-3	Full	90.3	5.0	4.7	0.801	90.6	5.1	4.3	0.807
	1-2-3	25% dec.	88.5	6.1	5.3	0.764	88.5	6.1	5.3	0.764
	1-2-3	50% dec.	85.8	7.8	6.4	0.709	86.1	7.5	6.4	0.715
	1-3-2	50% inc.	91.5	4.5	4.0	0.825	91.6	4.4	4.0	0.827
	1-3-2	Full	89.9	5.4	4.7	0.793	90.2	5.4	4.5	0.798
	1-3-2	25% dec.	88.2	6.4	5.4	0.758	89.1	6.0	4.9	0.775
	1-3-2	50% dec.	86.0	7.5	6.5	0.712	86.2	7.7	6.1	0.716
2	1-2-2	50% inc.	92.2	4.1	3.7	0.839	91.4	4.6	4.0	0.823
	1-2-2	Full	89.2	5.6	5.2	0.783	89.8	5.3	4.9	0.796
	1-2-2	25% dec.	89.2	5.9	4.9	0.778	88.9	6.0	5.1	0.772
	1-2-2	50% dec.	85.5	7.9	6.6	0.701	86.8	7.0	6.2	0.728
	1-3-3	50% inc.	91.9	4.5	3.5	0.834	91.9	4.5	3.6	0.833
	1-3-3	Full	89.7	5.8	4.4	0.789	90.4	5.4	4.3	0.802
	1-3-3	25% dec.	88.3	6.5	5.2	0.760	88.8	6.0	5.1	0.770
	1-3-3	50% dec.	86.3	7.6	6.1	0.720	86.0	7.7	6.2	0.713
	1-2-3	50% inc.	91.9	4.3	3.8	0.833	91.4	4.8	3.8	0.824
	1-2-3	Full	90.2	5.5	4.3	0.798	90.0	5.4	4.5	0.795
	1-2-3	25% dec.	88.7	6.3	5.0	0.767	88.6	6.1	5.3	0.766
	1-2-3	50% dec.	86.1	7.5	6.4	0.715	86.5	7.5	6.0	0.722
	1-3-2	50% inc.	91.9	4.3	3.7	0.834	91.4	4.9	3.7	0.823
	1-3-2	Full	89.8	5.6	4.6	0.791	90.5	5.3	4.1	0.805
	1-3-2	25% dec.	88.3	6.2	5.5	0.759	89.2	5.6	5.1	0.778
	1-3-2	50% dec.	86.3	7.4	6.3	0.719	86.7	7.4	5.8	0.728

Table 4.9. Decision Accuracy for DPI Routing at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.4	4.1	4.6	0.828	91.2	4.4	4.4	0.824
	1-2-2	Full	91.4	4.6	4.1	0.797	90.0	5.3	4.7	0.766
	1-2-2	25% dec.	88.0	5.7	6.3	0.761	88.8	5.6	5.7	0.776
	1-2-2	50% dec.	85.2	7.5	7.3	0.705	85.8	6.7	7.5	0.716
	1-3-3	50% inc.	91.1	4.4	4.5	0.823	91.4	4.1	4.5	0.828
	1-3-3	Full	89.4	5.1	5.4	0.789	89.9	5.0	5.1	0.799
	1-3-3	25% dec.	88.5	5.6	6.0	0.769	88.5	5.5	6.0	0.770
	1-3-3	50% dec.	85.8	6.8	7.4	0.716	86.1	6.8	7.1	0.721
	1-2-3	50% inc.	91.8	3.9	4.3	0.836	91.7	3.9	4.4	0.834
	1-2-3	Full	89.9	5.0	5.2	0.797	90.0	4.6	5.4	0.800
	1-2-3	25% dec.	89.0	5.1	5.9	0.780	88.4	5.5	6.2	0.767
	1-2-3	50% dec.	86.3	6.4	7.2	0.727	86.9	6.1	7.0	0.737
	1-3-2	50% inc.	91.6	4.0	4.4	0.833	91.5	4.2	4.3	0.830
	1-3-2	Full	89.7	5.0	5.4	0.793	90.2	4.9	5.0	0.804
	1-3-2	25% dec.	88.0	5.8	6.2	0.760	88.8	5.5	5.7	0.776
	1-3-2	50% dec.	86.1	6.6	7.3	0.722	86.5	6.3	7.1	0.731
2	1-2-2	50% inc.	91.3	4.3	4.4	0.826	91.0	4.4	4.6	0.821
	1-2-2	Full	91.2	5.0	3.8	0.795	89.9	5.7	4.4	0.763
	1-2-2	25% dec.	87.9	6.1	6.0	0.757	88.1	6.1	5.7	0.762
	1-2-2	50% dec.	85.9	7.4	6.7	0.717	85.9	6.9	7.3	0.717
	1-3-3	50% inc.	91.6	4.1	4.3	0.832	91.5	4.2	4.3	0.829
	1-3-3	Full	90.2	4.9	4.9	0.805	90.1	4.9	5.0	0.802
	1-3-3	25% dec.	88.4	5.9	5.7	0.768	88.9	5.8	5.3	0.777
	1-3-3	50% dec.	85.9	6.8	7.3	0.719	86.4	6.7	6.9	0.727
	1-2-3	50% inc.	91.6	3.9	4.5	0.833	91.1	4.7	4.2	0.822
	1-2-3	Full	90.2	4.9	4.9	0.804	90.2	4.6	5.2	0.803
	1-2-3	25% dec.	88.6	5.8	5.7	0.771	89.3	5.3	5.4	0.786
	1-2-3	50% dec.	86.1	7.0	6.9	0.723	86.5	6.8	6.6	0.730
	1-3-2	50% inc.	91.5	3.9	4.6	0.829	91.5	4.3	4.3	0.829
	1-3-2	Full	90.2	4.8	5.0	0.804	89.8	4.6	5.5	0.797
	1-3-2	25% dec.	88.5	5.8	5.7	0.769	89.0	5.4	5.7	0.780
	1-3-2	50% dec.	86.4	6.6	7.0	0.727	86.4	6.8	6.8	0.727

Table 4.10. Decision Accuracy for Proximity Routing at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.8	4.0	4.2	0.836	91.1	4.5	4.4	0.822
	1-2-2	Full	91.4	4.6	4.0	0.797	90.3	5.2	4.5	0.773
	1-2-2	25% dec.	87.9	6.0	6.0	0.759	88.4	5.7	5.8	0.769
	1-2-2	50% dec.	85.6	7.2	7.2	0.713	85.8	7.0	7.2	0.716
	1-3-3	50% inc.	91.8	4.2	4.1	0.835	91.3	4.4	4.3	0.826
	1-3-3	Full	89.8	5.1	5.1	0.795	89.8	5.2	5.0	0.796
	1-3-3	25% dec.	88.1	6.0	5.8	0.763	88.2	6.1	5.7	0.764
	1-3-3	50% dec.	85.9	6.9	7.2	0.718	85.9	7.1	7.0	0.717
	1-2-3	50% inc.	91.7	4.2	4.1	0.835	91.0	4.6	4.4	0.820
	1-2-3	Full	89.8	5.2	5.1	0.795	89.7	5.1	5.2	0.795
	1-2-3	25% dec.	87.8	6.2	6.1	0.755	88.2	6.1	5.7	0.764
	1-2-3	50% dec.	85.9	7.1	7.0	0.718	85.7	7.2	7.0	0.715
	1-3-2	50% inc.	91.5	4.4	4.1	0.831	91.3	4.4	4.2	0.826
	1-3-2	Full	89.7	5.1	5.2	0.794	90.0	5.0	5.0	0.800
	1-3-2	25% dec.	88.0	6.0	6.0	0.760	88.2	6.1	5.7	0.764
	1-3-2	50% dec.	86.0	6.9	7.0	0.721	85.8	7.1	7.1	0.716
2	1-2-2	50% inc.	91.8	4.2	4.0	0.837	91.2	4.5	4.4	0.823
	1-2-2	Full	91.5	4.8	3.7	0.800	90.3	5.5	4.2	0.773
	1-2-2	25% dec.	88.2	6.0	5.8	0.764	88.5	5.8	5.7	0.770
	1-2-2	50% dec.	85.6	7.4	7.0	0.713	86.1	7.1	6.9	0.721
	1-3-3	50% inc.	91.6	4.1	4.4	0.831	91.1	4.4	4.5	0.822
	1-3-3	Full	90.0	5.1	4.9	0.801	89.9	5.0	5.1	0.798
	1-3-3	25% dec.	88.4	5.6	6.0	0.768	88.6	5.6	5.8	0.773
	1-3-3	50% dec.	86.4	6.8	6.8	0.728	85.9	7.1	7.0	0.718
	1-2-3	50% inc.	91.7	4.2	4.2	0.833	91.3	4.4	4.3	0.826
	1-2-3	Full	89.7	5.1	5.2	0.794	90.1	5.0	4.9	0.801
	1-2-3	25% dec.	88.3	5.8	5.9	0.766	88.3	5.8	5.9	0.766
	1-2-3	50% dec.	86.2	7.1	6.7	0.724	86.2	7.0	6.8	0.725
	1-3-2	50% inc.	91.5	4.2	4.2	0.831	91.3	4.4	4.3	0.826
	1-3-2	Full	89.9	5.1	5.1	0.797	89.9	4.9	5.2	0.799
	1-3-2	25% dec.	88.1	5.8	6.0	0.762	88.6	5.6	5.8	0.772
	1-3-2	50% dec.	86.0	6.9	7.1	0.721	85.7	7.3	6.9	0.715

Table 4.11. Decision Accuracy for Number-Correct Routing at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.7	4.1	4.2	0.834	91.1	4.4	4.5	0.798
	1-2-2	Full	91.4	4.6	4.0	0.797	90.3	5.3	4.4	0.772
	1-2-2	25% dec.	87.6	6.1	6.3	0.751	88.4	5.7	5.9	0.768
	1-2-2	50% dec.	85.5	7.0	7.5	0.711	85.7	7.0	7.3	0.715
	1-3-3	50% inc.	91.5	4.1	4.4	0.831	91.3	4.3	4.4	0.826
	1-3-3	Full	89.5	5.2	5.3	0.789	89.8	5.1	5.0	0.797
	1-3-3	25% dec.	88.1	5.9	6.0	0.762	88.2	5.8	6.0	0.764
	1-3-3	50% dec.	85.7	7.0	7.3	0.714	86.0	6.9	7.0	0.721
	1-2-3	50% inc.	91.2	4.4	4.3	0.825	91.4	4.2	4.3	0.828
	1-2-3	Full	89.4	5.4	5.2	0.788	89.9	5.1	4.9	0.799
	1-2-3	25% dec.	87.8	6.1	6.1	0.756	88.1	6.1	5.9	0.761
	1-2-3	50% dec.	85.8	7.1	7.1	0.716	85.5	7.1	7.3	0.711
	1-3-2	50% inc.	91.5	4.2	4.3	0.829	91.1	4.5	4.4	0.822
	1-3-2	Full	89.3	5.3	5.4	0.786	89.9	5.2	4.9	0.798
	1-3-2	25% dec.	87.6	6.2	6.2	0.752	88.3	5.8	5.9	0.766
	1-3-2	50% dec.	86.0	6.9	7.1	0.719	85.7	7.3	7.0	0.714
2	1-2-2	50% inc.	91.6	4.2	4.2	0.831	91.0	4.6	4.4	0.821
	1-2-2	Full	91.5	4.8	3.7	0.801	90.4	5.5	4.2	0.774
	1-2-2	25% dec.	88.1	5.9	6.0	0.762	88.2	6.0	5.8	0.764
	1-2-2	50% dec.	85.9	7.0	7.1	0.718	86.0	6.8	7.2	0.719
	1-3-3	50% inc.	91.4	4.3	4.3	0.829	91.1	4.4	4.4	0.823
	1-3-3	Full	89.9	5.2	4.9	0.798	89.8	5.2	5.0	0.795
	1-3-3	25% dec.	88.3	5.8	5.9	0.765	88.4	5.8	5.8	0.769
	1-3-3	50% dec.	86.4	6.9	6.7	0.728	85.8	7.0	7.2	0.716
	1-2-3	50% inc.	91.9	4.2	3.9	0.838	91.1	4.5	4.4	0.822
	1-2-3	Full	90.1	5.0	4.9	0.802	89.9	5.0	5.1	0.798
	1-2-3	25% dec.	88.3	5.7	5.9	0.767	88.5	5.7	5.7	0.771
	1-2-3	50% dec.	86.0	7.3	6.7	0.719	85.9	7.1	7.0	0.717
	1-3-2	50% inc.	91.6	4.2	4.3	0.831	91.2	4.5	4.3	0.824
	1-3-2	Full	89.7	5.4	4.9	0.794	89.8	5.1	5.1	0.797
	1-3-2	25% dec.	87.9	5.9	6.1	0.759	88.6	5.6	5.8	0.771
	1-3-2	50% dec.	86.1	7.0	6.9	0.722	85.9	7.1	7.0	0.718

Table 4.12. Decision Accuracy for Random Routing at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree (%)	FP (%)	FN (%)	κ	Agree (%)	FP (%)	FN (%)	κ
1	1-2-2	50% inc.	91.1	4.4	4.4	0.823	91.0	4.4	4.5	0.821
	1-2-2	Full	91.3	4.9	3.8	0.796	89.9	5.6	4.5	0.764
	1-2-2	25% dec.	88.2	6.0	5.9	0.763	88.4	5.8	5.8	0.767
	1-2-2	50% dec.	85.2	7.5	7.3	0.704	86.1	7.2	6.7	0.722
	1-3-3	50% inc.	91.7	4.3	4.0	0.834	91.4	4.5	4.1	0.829
	1-3-3	Full	89.8	5.1	5.2	0.796	89.5	5.3	5.2	0.790
	1-3-3	25% dec.	87.6	6.2	6.2	0.752	88.0	6.1	5.9	0.760
	1-3-3	50% dec.	86.5	6.6	6.9	0.729	86.2	7.2	6.7	0.723
	1-2-3	50% inc.	91.2	4.5	4.3	0.825	91.2	4.7	4.1	0.823
	1-2-3	Full	89.6	5.4	5.0	0.791	89.7	5.4	5.0	0.794
	1-2-3	25% dec.	88.1	6.0	5.9	0.763	88.8	5.9	5.4	0.776
	1-2-3	50% dec.	86.2	7.2	6.6	0.723	85.7	7.5	6.9	0.713
	1-3-2	50% inc.	91.5	4.4	4.1	0.830	91.2	4.5	4.3	0.824
	1-3-2	Full	89.5	5.7	4.8	0.791	89.5	5.5	5.0	0.791
	1-3-2	25% dec.	87.7	6.1	6.2	0.754	88.4	6.0	5.6	0.768
	1-3-2	50% dec.	85.2	7.8	7.0	0.704	86.2	6.7	7.1	0.723
2	1-2-2	50% inc.	91.2	4.5	4.3	0.824	91.0	4.7	4.3	0.820
	1-2-2	Full	90.8	5.0	4.2	0.783	89.9	5.6	4.5	0.763
	1-2-2	25% dec.	88.3	6.0	5.7	0.766	88.0	5.8	6.1	0.761
	1-2-2	50% dec.	85.8	7.3	6.9	0.716	86.2	7.0	6.8	0.723
	1-3-3	50% inc.	91.5	4.4	4.1	0.829	91.1	4.5	4.4	0.823
	1-3-3	Full	89.6	5.4	5.0	0.791	89.6	5.3	5.1	0.793
	1-3-3	25% dec.	87.9	6.1	6.1	0.757	88.4	5.9	5.7	0.768
	1-3-3	50% dec.	85.4	7.6	7.0	0.707	85.7	7.3	7.0	0.713
	1-2-3	50% inc.	91.3	4.6	4.1	0.826	91.1	4.4	4.4	0.823
	1-2-3	Full	89.5	5.4	5.1	0.789	89.8	5.0	5.2	0.796
	1-2-3	25% dec.	88.5	5.9	5.6	0.770	88.3	5.9	5.7	0.767
	1-2-3	50% dec.	86.0	6.9	7.1	0.720	85.4	7.6	7.0	0.708
	1-3-2	50% inc.	91.5	4.3	4.2	0.831	90.8	5.0	4.2	0.817
	1-3-2	Full	89.4	5.3	5.3	0.789	89.9	5.2	4.9	0.797
	1-3-2	25% dec.	87.6	6.3	6.1	0.752	88.3	5.8	6.0	0.765
	1-3-2	50% dec.	86.7	6.9	6.4	0.734	85.6	7.3	7.1	0.713

Table 4.13. Decision Consistency for DPI Routing at Three Pass Rates

Pass Rate	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree	FP(%)	FN(%)	κ	Agree	FP(%)	FN(%)	κ
30%	1-2-2	50% inc.	89.8	5.5	4.7	0.762	89.4	5.2	5.4	0.752
	1-2-2	Full	88.5	6.1	5.4	0.731	87.8	6.4	5.8	0.716
	1-2-2	25% dec.	85.6	7.5	6.9	0.665	85.8	7.3	6.8	0.672
	1-2-2	50% dec.	83.1	8.6	8.3	0.610	82.9	8.9	8.2	0.606
	1-3-3	50% inc.	90.1	5.0	4.9	0.767	89.6	5.0	5.4	0.755
	1-3-3	Full	88.3	6.0	5.7	0.726	87.9	6.1	6.0	0.717
	1-3-3	25% dec.	86.6	7.0	6.5	0.688	86.2	7.1	6.8	0.678
	1-3-3	50% dec.	82.6	9.0	8.5	0.599	83.5	8.3	8.2	0.620
	1-2-3	50% inc.	90.0	5.0	5.0	0.766	89.8	5.2	5.1	0.760
	1-2-3	Full	88.6	5.9	5.6	0.733	87.2	6.5	6.3	0.703
	1-2-3	25% dec.	85.4	7.2	7.4	0.660	86.3	6.8	6.9	0.682
	1-2-3	50% dec.	82.8	8.9	8.3	0.602	82.9	8.5	8.6	0.607
	1-3-2	50% inc.	90.1	4.9	5.0	0.769	89.5	4.9	5.6	0.755
	1-3-2	Full	88.5	6.0	5.5	0.731	88.0	6.2	5.8	0.719
	1-3-2	25% dec.	85.9	7.4	6.7	0.672	85.7	7.2	7.0	0.670
	1-3-2	50% dec.	82.4	8.8	8.8	0.596	82.8	8.8	8.4	0.606
40%	1-2-2	50% inc.	88.7	5.9	5.4	0.768	88.1	6.4	5.5	0.756
	1-2-2	Full	85.7	7.4	6.9	0.706	86.5	7.1	6.4	0.722
	1-2-2	25% dec.	84.4	7.9	7.7	0.679	84.8	7.9	7.3	0.687
	1-2-2	50% dec.	80.8	9.8	9.4	0.606	81.4	9.5	9.1	0.618
	1-3-3	50% inc.	89.3	5.5	5.3	0.780	88.1	6.2	5.7	0.755
	1-3-3	Full	86.2	7.1	6.7	0.717	86.5	6.9	6.6	0.722
	1-3-3	25% dec.	83.9	8.3	7.8	0.668	84.6	7.8	7.6	0.684
	1-3-3	50% dec.	81.9	9.2	8.9	0.628	81.3	9.2	9.5	0.616
	1-2-3	50% inc.	88.9	5.6	5.5	0.771	88.4	6.2	5.5	0.760
	1-2-3	Full	86.4	7.3	6.3	0.720	86.9	6.9	6.3	0.730
	1-2-3	25% dec.	84.3	7.9	7.8	0.677	84.8	7.5	7.7	0.688
	1-2-3	50% dec.	81.3	9.9	8.9	0.616	81.4	9.6	9.0	0.618
	1-3-2	50% inc.	89.1	5.4	5.6	0.774	88.1	5.7	6.2	0.756
	1-3-2	Full	86.7	6.6	6.7	0.726	86.6	6.8	6.6	0.725
	1-3-2	25% dec.	84.3	8.1	7.7	0.677	84.9	7.4	7.7	0.690
	1-3-2	50% dec.	81.5	9.3	9.3	0.619	81.3	9.2	9.5	0.615
50%	1-2-2	50% inc.	88.0	6.2	5.8	0.760	87.7	6.0	6.3	0.754
	1-2-2	Full	85.8	7.6	6.7	0.716	85.9	7.1	7.0	0.719
	1-2-2	25% dec.	84.1	8.3	7.6	0.681	84.3	8.1	7.6	0.687
	1-2-2	50% dec.	80.2	10.2	9.6	0.603	81.1	9.7	9.2	0.622
	1-3-3	50% inc.	88.2	5.8	6.0	0.764	87.9	6.2	5.9	0.758
	1-3-3	Full	86.0	7.1	6.8	0.720	86.1	6.9	6.9	0.722
	1-3-3	25% dec.	84.0	8.3	7.6	0.680	84.7	8.1	7.2	0.694
	1-3-3	50% dec.	80.6	9.8	9.6	0.611	80.6	9.8	9.6	0.612
	1-2-3	50% inc.	88.3	5.8	5.9	0.767	87.6	6.7	5.6	0.753
	1-2-3	Full	86.3	6.9	6.8	0.726	86.2	7.0	6.8	0.725
	1-2-3	25% dec.	84.3	8.3	7.4	0.686	85.3	7.6	7.0	0.707
	1-2-3	50% dec.	80.7	10.0	9.2	0.615	81.5	9.8	8.7	0.630
	1-3-2	50% inc.	88.6	5.6	5.9	0.771	87.9	6.0	6.0	0.759
	1-3-2	Full	86.5	6.8	6.6	0.731	86.4	6.4	7.2	0.727
	1-3-2	25% dec.	83.8	8.4	7.8	0.676	84.9	7.5	7.6	0.698
	1-3-2	50% dec.	80.8	9.7	9.5	0.616	81.3	9.7	9.0	0.626

Table 4.14. Decision Consistency for Proximity Routing at Three Pass Rates

Pass Rate	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree	FP(%)	FN(%)	κ	Agree	FP(%)	FN(%)	κ
30%	1-2-2	50% inc.	90.2	5.2	4.6	0.769	89.4	5.3	5.3	0.752
	1-2-2	Full	88.0	6.4	5.6	0.719	88.0	6.3	5.7	0.719
	1-2-2	25% dec.	86.1	7.2	6.6	0.678	85.4	7.4	7.2	0.662
	1-2-2	50% dec.	82.7	8.9	8.5	0.600	83.2	8.8	8.1	0.613
	1-3-3	50% inc.	90.1	5.1	4.8	0.769	89.4	5.5	5.1	0.753
	1-3-3	Full	88.1	6.1	5.8	0.723	88.1	6.0	5.9	0.723
	1-3-3	25% dec.	86.3	7.1	6.6	0.680	85.5	7.3	7.3	0.662
	1-3-3	50% dec.	82.7	8.8	8.5	0.602	83.1	8.4	8.5	0.610
	1-2-3	50% inc.	90.1	5.2	4.7	0.768	89.6	5.4	5.0	0.757
	1-2-3	Full	87.6	6.3	6.1	0.710	87.9	6.0	6.1	0.718
	1-2-3	25% dec.	86.2	7.1	6.6	0.680	85.5	7.6	6.9	0.663
	1-2-3	50% dec.	82.7	8.6	8.7	0.603	83.3	8.6	8.1	0.615
	1-3-2	50% inc.	90.3	5.1	4.6	0.773	89.4	5.6	4.9	0.753
	1-3-2	Full	87.8	6.3	5.9	0.715	88.0	6.0	6.0	0.720
	1-3-2	25% dec.	86.2	7.2	6.6	0.680	85.7	7.1	7.2	0.670
	1-3-2	50% dec.	83.3	8.6	8.2	0.615	82.6	8.6	8.8	0.600
40%	1-2-2	50% inc.	88.3	5.7	6.0	0.760	88.3	5.6	6.1	0.760
	1-2-2	Full	86.4	6.8	6.9	0.720	86.3	6.7	7.1	0.718
	1-2-2	25% dec.	84.1	8.1	7.8	0.674	84.2	7.9	7.9	0.676
	1-2-2	50% dec.	80.7	9.8	9.5	0.605	80.6	9.5	9.9	0.602
	1-3-3	50% inc.	88.7	5.5	5.8	0.769	88.1	5.9	6.0	0.756
	1-3-3	Full	86.3	7.1	6.6	0.718	86.1	6.9	6.9	0.715
	1-3-3	25% dec.	84.3	7.8	8.0	0.677	84.3	7.8	7.9	0.679
	1-3-3	50% dec.	80.7	9.6	9.6	0.606	80.7	10.0	9.2	0.606
	1-2-3	50% inc.	88.3	5.9	5.9	0.759	88.4	5.8	5.8	0.761
	1-2-3	Full	86.1	6.9	7.0	0.714	85.7	7.0	7.3	0.707
	1-2-3	25% dec.	84.0	8.1	7.9	0.672	84.2	8.0	7.8	0.676
	1-2-3	50% dec.	80.0	10.0	10.0	0.590	81.0	9.5	9.4	0.612
	1-3-2	50% inc.	88.7	5.6	5.7	0.768	88.3	5.7	6.0	0.759
	1-3-2	Full	86.2	7.1	6.7	0.717	86.0	6.9	7.1	0.714
	1-3-2	25% dec.	84.1	8.0	7.8	0.675	84.3	8.0	7.7	0.678
	1-3-2	50% dec.	80.5	9.9	9.6	0.602	81.1	9.7	9.2	0.614
50%	1-2-2	50% inc.	88.4	6.0	5.7	0.768	91.2	4.5	4.4	0.823
	1-2-2	Full	85.3	7.4	7.3	0.705	86.3	6.9	6.8	0.725
	1-2-2	25% dec.	83.7	8.3	8.0	0.674	84.1	8.1	7.8	0.683
	1-2-2	50% dec.	79.9	10.2	9.9	0.597	80.1	10.1	9.7	0.603
	1-3-3	50% inc.	88.1	5.7	6.1	0.763	87.7	6.1	6.3	0.753
	1-3-3	Full	85.2	7.5	7.3	0.704	85.5	7.1	7.4	0.710
	1-3-3	25% dec.	83.8	7.8	8.4	0.677	83.7	7.9	8.4	0.675
	1-3-3	50% dec.	80.5	9.9	9.6	0.610	80.4	9.8	9.8	0.607
	1-2-3	50% inc.	87.8	6.0	6.2	0.756	87.3	6.2	6.5	0.746
	1-2-3	Full	85.4	7.2	7.4	0.708	85.8	7.2	7.0	0.716
	1-2-3	25% dec.	83.3	8.2	8.5	0.666	83.6	7.9	8.5	0.672
	1-2-3	50% dec.	80.7	9.7	9.5	0.615	80.3	9.8	9.8	0.606
	1-3-2	50% inc.	87.6	6.0	6.3	0.753	87.7	6.1	6.2	0.753
	1-3-2	Full	85.2	7.4	7.3	0.705	85.7	7.0	7.3	0.713
	1-3-2	25% dec.	84.0	7.9	8.1	0.681	83.6	7.9	8.4	0.672
	1-3-2	50% dec.	80.3	9.8	9.9	0.606	80.4	10.0	9.6	0.607

Table 4.15. Decision Consistency for Number-Correct Routing at Three Pass Rates

Pass Rate	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree	FP(%)	FN(%)	κ	Agree	FP(%)	FN(%)	κ
30%	1-2-2	50% inc.	90.2	5.3	4.6	0.769	89.4	5.3	5.3	0.752
	1-2-2	Full	87.9	6.5	5.7	0.716	87.9	6.3	5.8	0.718
	1-2-2	25% dec.	86.1	7.2	6.7	0.678	85.3	7.4	7.3	0.660
	1-2-2	50% dec.	82.7	8.8	8.6	0.600	83.0	8.9	8.1	0.610
	1-3-3	50% inc.	89.8	5.1	5.0	0.762	89.3	5.6	5.0	0.751
	1-3-3	Full	88.0	6.3	5.7	0.718	88.0	5.9	6.0	0.721
	1-3-3	25% dec.	86.2	7.2	6.6	0.679	85.4	7.4	7.2	0.660
	1-3-3	50% dec.	82.8	8.7	8.5	0.603	83.0	8.7	8.4	0.607
	1-2-3	50% inc.	90.3	5.1	4.6	0.772	89.7	5.3	5.0	0.758
	1-2-3	Full	87.5	6.5	6.0	0.708	87.8	6.1	6.1	0.715
	1-2-3	25% dec.	86.4	7.0	6.6	0.685	85.5	7.3	7.2	0.663
	1-2-3	50% dec.	82.8	8.7	8.5	0.605	83.3	8.5	8.2	0.615
	1-3-2	50% inc.	90.0	5.1	4.8	0.766	89.4	5.7	4.9	0.751
	1-3-2	Full	87.9	6.3	5.8	0.717	87.8	6.1	6.1	0.715
	1-3-2	25% dec.	86.2	7.2	6.6	0.679	85.7	7.2	7.1	0.669
	1-3-2	50% dec.	83.1	8.6	8.3	0.610	82.6	8.7	8.7	0.599
40%	1-2-2	50% inc.	88.2	5.7	6.1	0.757	88.4	5.7	5.9	0.761
	1-2-2	Full	86.2	6.8	7.0	0.716	86.4	6.7	6.9	0.721
	1-2-2	25% dec.	83.7	8.3	8.0	0.665	84.0	8.1	7.9	0.671
	1-2-2	50% dec.	80.3	9.7	9.9	0.597	80.5	9.6	9.9	0.601
	1-3-3	50% inc.	88.7	5.5	5.8	0.768	88.1	5.8	6.2	0.755
	1-3-3	Full	85.9	7.4	6.7	0.711	86.4	6.8	6.8	0.721
	1-3-3	25% dec.	84.0	8.0	8.0	0.672	84.6	7.6	7.8	0.684
	1-3-3	50% dec.	81.0	9.7	9.3	0.611	81.1	9.8	9.1	0.614
	1-2-3	50% inc.	88.2	5.8	6.0	0.758	88.1	5.8	6.1	0.755
	1-2-3	Full	85.8	6.8	7.4	0.708	85.7	6.9	7.4	0.706
	1-2-3	25% dec.	84.0	8.0	8.0	0.672	83.9	8.2	8.0	0.669
	1-2-3	50% dec.	80.0	10.4	9.7	0.589	80.6	9.7	9.6	0.604
	1-3-2	50% inc.	88.6	5.7	5.7	0.765	88.1	5.8	6.1	0.755
	1-3-2	Full	85.9	7.2	6.9	0.711	86.1	7.0	7.0	0.714
	1-3-2	25% dec.	83.8	8.3	8.0	0.667	84.5	7.8	7.7	0.582
	1-3-2	50% dec.	80.4	10.0	9.6	0.600	81.0	9.7	9.3	0.611
50%	1-2-2	50% inc.	88.3	5.9	5.8	0.766	87.1	6.6	6.3	0.742
	1-2-2	Full	85.3	7.7	7.0	0.706	85.9	7.3	6.9	0.718
	1-2-2	25% dec.	83.2	8.5	8.3	0.664	84.0	8.2	7.8	0.680
	1-2-2	50% dec.	80.0	10.2	9.8	0.599	79.9	10.0	10.1	0.597
	1-3-3	50% inc.	87.8	6.3	6.0	0.755	87.6	6.2	6.1	0.753
	1-3-3	Full	85.3	7.5	7.2	0.706	85.7	7.2	7.1	0.714
	1-3-3	25% dec.	83.8	8.1	8.1	0.677	83.6	8.3	8.1	0.672
	1-3-3	50% dec.	80.6	9.9	9.5	0.612	80.7	9.6	9.7	0.614
	1-2-3	50% inc.	88.0	6.1	5.9	0.760	87.5	6.3	6.2	0.750
	1-2-3	Full	85.5	7.2	7.3	0.710	86.0	6.8	7.2	0.720
	1-2-3	25% dec.	83.4	8.2	8.4	0.668	83.9	7.9	8.2	0.677
	1-2-3	50% dec.	80.8	10.0	9.2	0.616	79.8	10.2	10.0	0.596
	1-3-2	50% inc.	87.6	6.2	6.2	0.752	87.7	6.2	6.1	0.754
	1-3-2	Full	85.4	7.6	7.0	0.707	85.5	7.1	7.4	0.709
	1-3-2	25% dec.	83.3	8.3	8.4	0.666	83.9	8.0	8.1	0.678
	1-3-2	50% dec.	80.3	10.0	9.7	0.606	80.1	9.9	10.0	0.602

Table 4.16. Decision Consistency for Random Routing at Three Pass Rates

Pass Rate	Design	TIF Level	Equal Information				1/2-1/4-1/4 Information			
			Agree	FP(%)	FN(%)	κ	Agree	FP(%)	FN(%)	κ
30%	1-2-2	50% inc.	89.1	5.4	5.5	0.745	89.8	5.0	5.2	0.762
	1-2-2	Full	87.7	6.3	6.1	0.713	87.3	6.2	6.5	0.705
	1-2-2	25% dec.	85.5	7.2	7.3	0.663	85.4	7.1	7.5	0.663
	1-2-2	50% dec.	82.0	9.0	9.0	0.585	82.8	8.8	8.4	0.607
	1-3-3	50% inc.	89.1	5.4	5.5	0.745	89.0	5.5	5.5	0.742
	1-3-3	Full	87.8	6.5	5.7	0.715	87.8	5.8	6.4	0.714
	1-3-3	25% dec.	85.5	7.0	7.5	0.663	85.8	7.0	7.2	0.672
	1-3-3	50% dec.	82.9	8.4	8.6	0.610	82.6	8.9	8.5	0.602
	1-2-3	50% inc.	89.5	5.4	5.1	0.755	89.5	5.0	5.5	0.754
	1-2-3	Full	87.2	6.5	6.3	0.705	87.6	6.1	6.3	0.713
	1-2-3	25% dec.	85.0	7.3	7.7	0.653	85.4	7.7	6.9	0.662
	1-2-3	50% dec.	81.7	9.0	9.3	0.582	83.1	8.2	8.8	0.611
	1-3-2	50% inc.	89.6	5.2	5.2	0.756	90.1	4.8	5.1	0.767
	1-3-2	Full	87.4	6.3	6.4	0.709	87.5	6.1	6.4	0.709
	1-3-2	25% dec.	85.7	7.1	7.2	0.668	86.0	7.1	7.0	0.676
	1-3-2	50% dec.	82.6	8.8	8.6	0.602	82.6	8.9	8.4	0.600
40%	1-2-2	50% inc.	89.2	5.3	5.5	0.777	88.0	6.0	6.0	0.753
	1-2-2	Full	85.6	7.2	7.1	0.705	86.3	6.9	6.8	0.719
	1-2-2	25% dec.	84.3	7.9	7.8	0.678	84.1	8.1	7.8	0.674
	1-2-2	50% dec.	80.7	9.6	9.7	0.605	81.3	9.5	9.2	0.617
	1-3-3	50% inc.	88.2	6.0	5.8	0.758	88.0	6.0	6.0	0.754
	1-3-3	Full	85.9	7.3	6.8	0.711	86.3	7.0	6.6	0.720
	1-3-3	25% dec.	83.9	8.0	8.1	0.671	84.5	7.8	7.7	0.681
	1-3-3	50% dec.	80.3	10.1	9.6	0.597	81.0	9.5	9.5	0.611
	1-2-3	50% inc.	88.2	5.8	6.0	0.758	88.5	5.9	5.6	0.765
	1-2-3	Full	85.5	7.7	6.9	0.701	86.2	7.0	6.8	0.717
	1-2-3	25% dec.	83.9	8.3	7.8	0.670	84.1	7.9	7.9	0.674
	1-2-3	50% dec.	80.8	9.5	9.8	0.606	81.1	9.6	9.3	0.614
	1-3-2	50% inc.	88.6	5.8	5.6	0.765	88.5	6.1	5.4	0.764
	1-3-2	Full	85.6	7.3	7.1	0.704	86.3	7.0	6.7	0.719
	1-3-2	25% dec.	84.0	7.9	8.1	0.672	84.5	7.5	8.1	0.681
	1-3-2	50% dec.	80.6	9.7	9.7	0.603	80.9	9.6	9.5	0.609
50%	1-2-2	50% inc.	87.6	6.3	6.0	0.753	87.8	6.4	5.9	0.756
	1-2-2	Full	84.9	7.7	7.5	0.697	85.6	7.2	7.2	0.712
	1-2-2	25% dec.	83.4	8.4	8.2	0.668	83.7	8.0	8.3	0.674
	1-2-2	50% dec.	80.1	10.0	9.8	0.603	80.4	9.6	9.9	0.609
	1-3-3	50% inc.	88.0	6.0	6.0	0.761	87.9	5.9	6.2	0.758
	1-3-3	Full	85.3	7.6	7.1	0.706	85.9	7.0	7.0	0.719
	1-3-3	25% dec.	83.4	8.4	8.3	0.667	84.1	7.9	8.0	0.682
	1-3-3	50% dec.	80.3	10.3	9.4	0.606	80.5	9.7	9.8	0.610
	1-2-3	50% inc.	88.0	6.1	5.9	0.759	87.8	5.8	6.4	0.756
	1-2-3	Full	85.8	7.1	7.1	0.716	85.8	6.8	7.4	0.716
	1-2-3	25% dec.	83.5	8.4	8.1	0.670	84.4	7.6	7.9	0.688
	1-2-3	50% dec.	80.4	9.3	10.2	0.608	80.1	9.9	10.0	0.601
	1-3-2	50% inc.	88.3	5.8	6.0	0.765	87.6	6.5	6.0	0.751
	1-3-2	Full	85.2	7.0	7.9	0.703	86.0	6.9	7.1	0.709
	1-3-2	25% dec.	82.9	8.7	8.4	0.657	84.0	7.7	8.3	0.680
	1-3-2	50% dec.	81.1	9.3	9.7	0.621	80.1	10.2	9.7	0.603

Table 4.17. Correlations Between True and Estimated Abilities at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI $r_{t \cdot e_{final}}$	Prox. $r_{t \cdot e_{final}}$	NC $r_{t \cdot e_{final}}$	Ran. $r_{t \cdot e_{final}}$	DPI $r_{t \cdot e_{final}}$	Prox. $r_{t \cdot e_{final}}$	NC $r_{t \cdot e_{final}}$	Ran. $r_{t \cdot e_{final}}$
1	1-2-2	50% inc.	0.960	0.960	0.960	0.953	0.959	0.958	0.957	0.955
	1-2-2	Full	0.944	0.946	0.946	0.942	0.946	0.943	0.943	0.940
	1-2-2	25% dec.	0.933	0.931	0.931	0.929	0.930	0.929	0.929	0.927
	1-2-2	50% dec.	0.904	0.904	0.904	0.898	0.903	0.903	0.902	0.900
	1-3-3	50% inc.	0.961	0.959	0.959	0.955	0.958	0.957	0.956	0.952
	1-3-3	Full	0.943	0.945	0.945	0.942	0.947	0.945	0.945	0.940
	1-3-3	25% dec.	0.934	0.932	0.932	0.927	0.931	0.929	0.929	0.927
	1-3-3	50% dec.	0.907	0.904	0.905	0.899	0.907	0.901	0.901	0.900
	1-2-3	50% inc.	0.959	0.959	0.959	0.955	0.959	0.957	0.957	0.954
	1-2-3	Full	0.945	0.946	0.946	0.943	0.947	0.944	0.943	0.941
	1-2-3	25% dec.	0.933	0.932	0.932	0.926	0.932	0.927	0.928	0.926
	1-2-3	50% dec.	0.910	0.905	0.905	0.900	0.905	0.900	0.900	0.902
	1-3-2	50% inc.	0.960	0.959	0.959	0.956	0.959	0.956	0.956	0.952
	1-3-2	Full	0.944	0.946	0.945	0.943	0.948	0.944	0.944	0.940
	1-3-2	25% dec.	0.933	0.931	0.931	0.927	0.932	0.929	0.929	0.929
	1-3-2	50% dec.	0.906	0.905	0.906	0.900	0.906	0.901	0.900	0.899
2	1-2-2	50% inc.	0.960	0.959	0.959	0.953	0.958	0.959	0.959	0.953
	1-2-2	Full	0.949	0.947	0.947	0.941	0.944	0.945	0.946	0.941
	1-2-2	25% dec.	0.934	0.932	0.932	0.927	0.932	0.929	0.929	0.927
	1-2-2	50% dec.	0.908	0.905	0.905	0.903	0.905	0.903	0.902	0.902
	1-3-3	50% inc.	0.961	0.958	0.959	0.954	0.958	0.957	0.957	0.952
	1-3-3	Full	0.948	0.948	0.948	0.941	0.947	0.946	0.946	0.941
	1-3-3	25% dec.	0.934	0.934	0.934	0.928	0.932	0.929	0.929	0.926
	1-3-3	50% dec.	0.906	0.903	0.904	0.902	0.906	0.904	0.903	0.900
	1-2-3	50% inc.	0.961	0.960	0.960	0.956	0.958	0.957	0.957	0.953
	1-2-3	Full	0.947	0.947	0.947	0.942	0.946	0.946	0.946	0.940
	1-2-3	25% dec.	0.935	0.934	0.934	0.928	0.929	0.928	0.929	0.928
	1-2-3	50% dec.	0.909	0.905	0.905	0.900	0.907	0.904	0.903	0.902
	1-3-2	50% inc.	0.961	0.959	0.959	0.955	0.958	0.957	0.957	0.954
	1-3-2	Full	0.947	0.947	0.947	0.940	0.946	0.946	0.945	0.941
	1-3-2	25% dec.	0.934	0.934	0.934	0.926	0.931	0.929	0.929	0.926
	1-3-2	50% dec.	0.904	0.905	0.906	0.900	0.906	0.903	0.902	0.903

Table 4.18. Correlations Between True and Estimated Abilities at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI	Prox.	NC	Ran.	DPI	Prox.	NC	Ran.
			$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$	$r_{t \cdot e_{final}}$
1	1-2-2	50% inc.	0.960	0.957	0.957	0.954	0.958	0.957	0.957	0.952
	1-2-2	Full	0.946	0.945	0.945	0.941	0.947	0.943	0.943	0.938
	1-2-2	25% dec.	0.934	0.932	0.932	0.928	0.932	0.930	0.930	0.927
	1-2-2	50% dec.	0.902	0.904	0.903	0.902	0.903	0.903	0.902	0.899
	1-3-3	50% inc.	0.960	0.958	0.959	0.954	0.958	0.955	0.955	0.952
	1-3-3	Full	0.947	0.947	0.947	0.942	0.947	0.944	0.944	0.942
	1-3-3	25% dec.	0.933	0.932	0.932	0.926	0.932	0.929	0.929	0.929
	1-3-3	50% dec.	0.906	0.904	0.903	0.898	0.905	0.902	0.901	0.899
	1-2-3	50% inc.	0.960	0.957	0.959	0.954	0.958	0.956	0.956	0.953
	1-2-3	Full	0.948	0.946	0.946	0.940	0.947	0.944	0.943	0.941
	1-2-3	25% dec.	0.935	0.933	0.933	0.926	0.933	0.928	0.928	0.929
	1-2-3	50% dec.	0.907	0.904	0.903	0.899	0.904	0.902	0.901	0.898
	1-3-2	50% inc.	0.961	0.959	0.959	0.954	0.958	0.956	0.956	0.952
	1-3-2	Full	0.947	0.946	0.946	0.940	0.947	0.943	0.943	0.941
	1-3-2	25% dec.	0.932	0.932	0.932	0.926	0.932	0.929	0.929	0.928
	1-3-2	50% dec.	0.906	0.904	0.903	0.898	0.903	0.903	0.902	0.901
2	1-2-2	50% inc.	0.959	0.958	0.958	0.953	0.957	0.957	0.957	0.953
	1-2-2	Full	0.946	0.946	0.946	0.940	0.945	0.945	0.944	0.940
	1-2-2	25% dec.	0.934	0.932	0.932	0.927	0.933	0.930	0.930	0.930
	1-2-2	50% dec.	0.906	0.904	0.904	0.897	0.905	0.902	0.902	0.902
	1-3-3	50% inc.	0.960	0.960	0.960	0.953	0.958	0.957	0.957	0.951
	1-3-3	Full	0.947	0.947	0.946	0.941	0.945	0.946	0.946	0.940
	1-3-3	25% dec.	0.933	0.933	0.933	0.927	0.932	0.931	0.931	0.928
	1-3-3	50% dec.	0.907	0.902	0.902	0.899	0.904	0.900	0.900	0.899
	1-2-3	50% inc.	0.961	0.958	0.959	0.954	0.957	0.958	0.958	0.952
	1-2-3	Full	0.948	0.947	0.947	0.941	0.947	0.946	0.946	0.939
	1-2-3	25% dec.	0.934	0.933	0.933	0.928	0.933	0.930	0.930	0.928
	1-2-3	50% dec.	0.908	0.904	0.905	0.901	0.902	0.901	0.900	0.899
	1-3-2	50% inc.	0.961	0.960	0.960	0.954	0.959	0.958	0.958	0.954
	1-3-2	Full	0.949	0.947	0.947	0.940	0.945	0.945	0.945	0.940
	1-3-2	25% dec.	0.934	0.933	0.932	0.926	0.931	0.930	0.930	0.929
	1-3-2	50% dec.	0.905	0.903	0.903	0.900	0.904	0.901	0.901	0.902

Table 4.19. Correlations Between True and Estimated Abilities at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI	Prox.	NC	Ran.	DPI	Prox.	NC	Ran.
			$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$	$r_{t-e_{final}}$
1	1-2-2	50% inc.	0.959	0.958	0.957	0.952	0.956	0.954	0.954	0.949
	1-2-2	Full	0.945	0.944	0.944	0.938	0.944	0.942	0.941	0.940
	1-2-2	25% dec.	0.932	0.929	0.929	0.922	0.930	0.926	0.926	0.926
	1-2-2	50% dec.	0.901	0.899	0.899	0.894	0.899	0.897	0.898	0.895
	1-3-3	50% inc.	0.958	0.958	0.957	0.953	0.956	0.955	0.955	0.950
	1-3-3	Full	0.947	0.943	0.943	0.937	0.945	0.942	0.942	0.938
	1-3-3	25% dec.	0.932	0.930	0.930	0.924	0.931	0.927	0.927	0.925
	1-3-3	50% dec.	0.904	0.902	0.903	0.896	0.899	0.898	0.897	0.896
	1-2-3	50% inc.	0.958	0.958	0.957	0.953	0.957	0.955	0.955	0.950
	1-2-3	Full	0.946	0.945	0.944	0.940	0.944	0.942	0.942	0.938
	1-2-3	25% dec.	0.933	0.930	0.930	0.924	0.928	0.927	0.926	0.926
	1-2-3	50% dec.	0.902	0.901	0.901	0.895	0.902	0.897	0.897	0.897
	1-3-2	50% inc.	0.959	0.957	0.957	0.953	0.956	0.956	0.956	0.951
	1-3-2	Full	0.945	0.944	0.944	0.938	0.945	0.943	0.942	0.939
	1-3-2	25% dec.	0.931	0.931	0.930	0.922	0.930	0.926	0.926	0.923
	1-3-2	50% dec.	0.903	0.903	0.903	0.894	0.902	0.899	0.898	0.899
2	1-2-2	50% inc.	0.960	0.958	0.958	0.952	0.955	0.955	0.955	0.950
	1-2-2	Full	0.944	0.945	0.945	0.939	0.944	0.942	0.942	0.941
	1-2-2	25% dec.	0.931	0.930	0.929	0.924	0.929	0.928	0.928	0.925
	1-2-2	50% dec.	0.900	0.901	0.900	0.896	0.899	0.897	0.896	0.896
	1-3-3	50% inc.	0.958	0.959	0.958	0.952	0.957	0.956	0.956	0.949
	1-3-3	Full	0.947	0.945	0.945	0.940	0.944	0.944	0.944	0.940
	1-3-3	25% dec.	0.932	0.931	0.931	0.925	0.932	0.929	0.928	0.925
	1-3-3	50% dec.	0.905	0.905	0.905	0.898	0.904	0.897	0.896	0.897
	1-2-3	50% inc.	0.960	0.958	0.958	0.954	0.957	0.957	0.957	0.950
	1-2-3	Full	0.948	0.945	0.945	0.939	0.944	0.944	0.943	0.939
	1-2-3	25% dec.	0.933	0.931	0.931	0.921	0.931	0.928	0.928	0.926
	1-2-3	50% dec.	0.904	0.904	0.903	0.899	0.901	0.896	0.896	0.895
	1-3-2	50% inc.	0.957	0.958	0.957	0.952	0.957	0.956	0.956	0.952
	1-3-2	Full	0.947	0.944	0.944	0.939	0.943	0.945	0.944	0.939
	1-3-2	25% dec.	0.932	0.931	0.930	0.922	0.931	0.928	0.928	0.922
	1-3-2	50% dec.	0.906	0.903	0.903	0.896	0.902	0.896	0.895	0.896

Table 4.20. Overall Root Mean Square Errors at 30% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE	DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE
1	1-2-2	50% inc.	0.30	0.30	0.30	0.33	0.30	0.31	0.31	0.33
	1-2-2	Full	0.36	0.35	0.35	0.37	0.35	0.36	0.36	0.37
	1-2-2	25% dec.	0.39	0.39	0.39	0.41	0.41	0.41	0.41	0.42
	1-2-2	50% dec.	0.48	0.48	0.48	0.50	0.49	0.48	0.49	0.50
	1-3-3	50% inc.	0.30	0.30	0.30	0.32	0.31	0.31	0.31	0.33
	1-3-3	Full	0.36	0.35	0.35	0.37	0.35	0.35	0.35	0.38
	1-3-3	25% dec.	0.39	0.39	0.39	0.41	0.40	0.41	0.41	0.42
	1-3-3	50% dec.	0.47	0.48	0.47	0.49	0.48	0.49	0.49	0.49
	1-2-3	50% inc.	0.30	0.30	0.30	0.32	0.31	0.31	0.31	0.32
	1-2-3	Full	0.36	0.35	0.35	0.36	0.35	0.36	0.36	0.37
	1-2-3	25% dec.	0.39	0.39	0.39	0.42	0.40	0.42	0.41	0.42
	1-2-3	50% dec.	0.47	0.48	0.47	0.49	0.49	0.49	0.49	0.49
	1-3-2	50% inc.	0.30	0.30	0.30	0.32	0.30	0.31	0.31	0.34
	1-3-2	Full	0.35	0.35	0.35	0.36	0.35	0.35	0.35	0.38
	1-3-2	25% dec.	0.39	0.39	0.39	0.41	0.40	0.41	0.41	0.41
	1-3-2	50% dec.	0.47	0.47	0.47	0.49	0.48	0.49	0.49	0.50
2	1-2-2	50% inc.	0.30	0.30	0.30	0.33	0.31	0.30	0.30	0.33
	1-2-2	Full	0.34	0.34	0.34	0.37	0.36	0.35	0.35	0.37
	1-2-2	25% dec.	0.39	0.39	0.39	0.42	0.40	0.41	0.41	0.41
	1-2-2	50% dec.	0.47	0.47	0.47	0.49	0.48	0.49	0.49	0.49
	1-3-3	50% inc.	0.30	0.30	0.30	0.32	0.31	0.31	0.31	0.34
	1-3-3	Full	0.34	0.34	0.34	0.37	0.35	0.35	0.35	0.37
	1-3-3	25% dec.	0.39	0.39	0.39	0.41	0.39	0.41	0.41	0.42
	1-3-3	50% dec.	0.48	0.48	0.48	0.50	0.48	0.48	0.48	0.50
	1-2-3	50% inc.	0.29	0.30	0.30	0.32	0.31	0.31	0.31	0.33
	1-2-3	Full	0.34	0.35	0.35	0.37	0.35	0.35	0.35	0.38
	1-2-3	25% dec.	0.39	0.39	0.39	0.41	0.41	0.41	0.41	0.41
	1-2-3	50% dec.	0.47	0.48	0.48	0.50	0.47	0.48	0.49	0.49
	1-3-2	50% inc.	0.30	0.30	0.30	0.32	0.31	0.31	0.31	0.33
	1-3-2	Full	0.35	0.35	0.35	0.37	0.35	0.35	0.35	0.37
	1-3-2	25% dec.	0.39	0.39	0.39	0.41	0.40	0.41	0.41	0.42
	1-3-2	50% dec.	0.48	0.48	0.47	0.50	0.48	0.49	0.49	0.49

Table 4.21. Overall Root Mean Square Errors at 40% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE	DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE
1	1-2-2	50% inc.	0.30	0.31	0.31	0.32	0.31	0.31	0.31	0.33
	1-2-2	Full	0.35	0.35	0.35	0.37	0.35	0.36	0.36	0.38
	1-2-2	25% dec.	0.39	0.39	0.39	0.42	0.39	0.40	0.40	0.41
	1-2-2	50% dec.	0.49	0.48	0.48	0.49	0.49	0.48	0.48	0.50
	1-3-3	50% inc.	0.30	0.30	0.30	0.32	0.31	0.32	0.32	0.33
	1-3-3	Full	0.35	0.34	0.34	0.37	0.35	0.36	0.35	0.37
	1-3-3	25% dec.	0.39	0.39	0.39	0.42	0.40	0.40	0.40	0.41
	1-3-3	50% dec.	0.48	0.48	0.48	0.50	0.49	0.49	0.49	0.50
	1-2-3	50% inc.	0.30	0.31	0.30	0.32	0.31	0.31	0.31	0.33
	1-2-3	Full	0.34	0.35	0.35	0.37	0.35	0.36	0.36	0.38
	1-2-3	25% dec.	0.39	0.39	0.39	0.42	0.40	0.41	0.41	0.41
	1-2-3	50% dec.	0.47	0.48	0.48	0.49	0.49	0.49	0.49	0.50
	1-3-2	50% inc.	0.30	0.30	0.30	0.33	0.31	0.31	0.31	0.34
	1-3-2	Full	0.34	0.35	0.35	0.37	0.35	0.36	0.36	0.37
	1-3-2	25% dec.	0.40	0.39	0.39	0.42	0.40	0.41	0.40	0.41
	1-3-2	50% dec.	0.48	0.48	0.48	0.50	0.49	0.49	0.49	0.50
2	1-2-2	50% inc.	0.30	0.31	0.31	0.33	0.31	0.31	0.31	0.33
	1-2-2	Full	0.35	0.35	0.35	0.37	0.35	0.36	0.36	0.38
	1-2-2	25% dec.	0.39	0.39	0.39	0.42	0.39	0.40	0.40	0.41
	1-2-2	50% dec.	0.48	0.48	0.48	0.50	0.49	0.49	0.49	0.49
	1-3-3	50% inc.	0.29	0.30	0.30	0.33	0.30	0.31	0.31	0.34
	1-3-3	Full	0.34	0.35	0.35	0.37	0.35	0.35	0.35	0.37
	1-3-3	25% dec.	0.39	0.39	0.39	0.42	0.40	0.40	0.40	0.41
	1-3-3	50% dec.	0.47	0.49	0.49	0.50	0.48	0.49	0.49	0.50
	1-2-3	50% inc.	0.29	0.31	0.30	0.33	0.31	0.31	0.31	0.33
	1-2-3	Full	0.34	0.34	0.34	0.37	0.34	0.35	0.35	0.38
	1-2-3	25% dec.	0.39	0.39	0.39	0.41	0.39	0.40	0.40	0.41
	1-2-3	50% dec.	0.47	0.48	0.48	0.49	0.49	0.49	0.49	0.50
	1-3-2	50% inc.	0.29	0.30	0.30	0.32	0.30	0.31	0.31	0.33
	1-3-2	Full	0.34	0.34	0.34	0.37	0.35	0.35	0.35	0.38
	1-3-2	25% dec.	0.39	0.39	0.39	0.41	0.40	0.40	0.40	0.41
	1-3-2	50% dec.	0.48	0.48	0.48	0.49	0.48	0.66	0.49	0.49

Table 4.22. Overall Root Mean Square Errors at 50% Passing

Rep.	Design	TIF Level	Equal Information				1/2-1/4-1/2 Information			
			DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE	DPI RMSE	Prox. RMSE	NC RMSE	Ran. RMSE
1	1-2-2	50% inc.	0.30	0.31	0.31	0.33	0.32	0.32	0.32	0.34
	1-2-2	Full	0.35	0.35	0.36	0.38	0.36	0.36	0.36	0.37
	1-2-2	25% dec.	0.39	0.40	0.40	0.43	0.41	0.42	0.42	0.42
	1-2-2	50% dec.	0.49	0.49	0.49	0.51	0.50	0.50	0.50	0.51
	1-3-3	50% inc.	0.31	0.31	0.31	0.33	0.32	0.32	0.32	0.34
	1-3-3	Full	0.35	0.36	0.36	0.38	0.36	0.36	0.36	0.38
	1-3-3	25% dec.	0.40	0.40	0.40	0.42	0.41	0.41	0.41	0.43
	1-3-3	50% dec.	0.49	0.48	0.48	0.51	0.50	0.50	0.50	0.51
	1-2-3	50% inc.	0.31	0.31	0.31	0.33	0.31	0.31	0.31	0.34
	1-2-3	Full	0.35	0.35	0.35	0.37	0.36	0.36	0.36	0.38
	1-2-3	25% dec.	0.39	0.40	0.40	0.42	0.42	0.41	0.42	0.42
	1-2-3	50% dec.	0.49	0.49	0.49	0.50	0.50	0.50	0.50	0.51
	1-3-2	50% inc.	0.31	0.31	0.31	0.33	0.32	0.31	0.31	0.34
	1-3-2	Full	0.36	0.35	0.35	0.38	0.35	0.36	0.36	0.38
	1-3-2	25% dec.	0.40	0.40	0.40	0.43	0.41	0.42	0.42	0.43
	1-3-2	50% dec.	0.49	0.48	0.48	0.51	0.49	0.50	0.50	0.50
2	1-2-2	50% inc.	0.30	0.31	0.31	0.34	0.32	0.32	0.32	0.34
	1-2-2	Full	0.36	0.35	0.35	0.38	0.36	0.37	0.36	0.37
	1-2-2	25% dec.	0.40	0.40	0.40	0.43	0.41	0.41	0.41	0.42
	1-2-2	50% dec.	0.50	0.49	0.49	0.51	0.50	0.50	0.51	0.51
	1-3-3	50% inc.	0.31	0.30	0.31	0.34	0.31	0.32	0.31	0.34
	1-3-3	Full	0.35	0.35	0.35	0.38	0.35	0.36	0.36	0.38
	1-3-3	25% dec.	0.39	0.40	0.40	0.42	0.40	0.41	0.41	0.42
	1-3-3	50% dec.	0.48	0.48	0.48	0.51	0.49	0.50	0.50	0.51
	1-2-3	50% inc.	0.30	0.31	0.31	0.32	0.31	0.31	0.31	0.34
	1-2-3	Full	0.34	0.35	0.35	0.38	0.36	0.36	0.36	0.38
	1-2-3	25% dec.	0.39	0.39	0.40	0.43	0.40	0.41	0.41	0.42
	1-2-3	50% dec.	0.48	0.48	0.48	0.51	0.49	0.51	0.51	0.52
	1-3-2	50% inc.	0.31	0.31	0.31	0.34	0.31	0.31	0.31	0.34
	1-3-2	Full	0.35	0.35	0.35	0.38	0.36	0.36	0.36	0.38
	1-3-2	25% dec.	0.39	0.40	0.40	0.43	0.40	0.41	0.41	0.43
	1-3-2	50% dec.	0.48	0.48	0.48	0.51	0.49	0.50	0.51	0.50

Table 4.23. Routing Path Frequencies in 1-2-2 Design with Four Routing Strategies

Division of Information	TIF Level	Module			Routing Strategy			
		s1	s2	s3	DPI	Proximity	NC	Random
Equal Information Across Stages	50% Increase	1	1	1	45.5%	36.3%	38.9%	25.2%
		1	1	2	4.5%	12.2%	7.9%	24.7%
		1	2	1	4.5%	7.8%	4.7%	25.2%
		1	2	2	45.5%	43.7%	47.5%	24.9%
	Full	1	1	1	44.6%	33.8%	37.5%	24.9%
		1	1	2	5.4%	10.9%	8.6%	25.1%
		1	2	1	5.4%	11.8%	4.4%	24.6%
		1	2	2	44.6%	43.5%	49.5%	25.4%
	25% Decrease	1	1	1	43.9%	30.8%	39.4%	25.2%
		1	1	2	6.1%	12.0%	8.7%	25.0%
		1	2	1	6.1%	14.0%	3.3%	24.8%
		1	2	2	43.9%	43.1%	48.6%	25.0%
	50% Decrease	1	1	1	42.7%	29.0%	41.2%	24.7%
		1	1	2	7.3%	13.9%	8.7%	25.2%
		1	2	1	7.3%	14.9%	3.4%	25.0%
		1	2	2	42.7%	42.2%	46.7%	25.1%
1/2-1/4-1/4 Information Across Stages	50% Increase	1	1	1	46.6%	42.8%	34.1%	24.9%
		1	1	2	3.4%	8.6%	12.3%	25.0%
		1	2	1	3.4%	5.7%	9.1%	25.1%
		1	2	2	46.6%	42.9%	44.5%	25.0%
	Full	1	1	1	46.4%	43.0%	34.6%	24.9%
		1	1	2	3.6%	8.9%	13.8%	25.1%
		1	2	1	3.6%	6.3%	8.8%	25.0%
		1	2	2	46.4%	42.8%	42.7%	25.0%
	25% Decrease	1	1	1	45.8%	39.8%	32.1%	24.7%
		1	1	2	4.2%	9.8%	12.4%	25.2%
		1	2	1	4.2%	8.0%	12.6%	25.1%
		1	2	2	45.8%	42.4%	43.0%	24.9%
	50% Decrease	1	1	1	45.0%	38.5%	28.7%	24.9%
		1	1	2	5.0%	9.8%	14.1%	24.8%
		1	2	1	5.0%	9.0%	15.5%	25.2%
		1	2	2	45.0%	43.8%	41.7%	25.1%

Table 4.24. Routing Path Frequencies in 1-3-3 Design with Four Routing Strategies

Division of Information	TIF Level	Module			Routing Strategy			
		s1	s2	s3	DPI	Proximity	NC	Random
Equal Information Across Stages	50% Increase	1	1	1	29.2%	22.1%	29.2%	14.7%
		1	1	2	4.1%	6.6%	3.7%	15.0%
		1	2	1	4.1%	1.3%	6.4%	14.1%
		1	2	2	24.8%	27.1%	13.6%	15.1%
		1	2	3	4.4%	5.6%	9.8%	14.0%
		1	3	2	4.4%	2.8%	3.4%	14.5%
		1	3	1	29.0%	33.6%	34.1%	13.7%
	Full	1	1	1	28.4%	18.9%	25.9%	13.5%
		1	1	2	4.9%	8.1%	5.6%	14.8%
		1	2	1	4.9%	1.5%	6.7%	14.3%
		1	2	2	23.5%	28.5%	12.4%	15.1%
		1	2	3	5.0%	8.8%	9.4%	14.3%
		1	3	2	4.9%	2.3%	5.7%	14.6%
		1	3	1	28.4%	30.6%	34.3%	15.5%
	25% Decrease	1	1	1	27.7%	20.4%	23.8%	14.6%
		1	1	2	5.6%	6.8%	6.7%	13.7%
		1	2	1	5.5%	3.0%	8.3%	14.2%
		1	2	2	22.2%	31.7%	12.4%	14.1%
		1	2	3	5.6%	6.8%	8.1%	14.2%
		1	3	2	5.6%	4.9%	5.8%	14.7%
		1	3	1	27.7%	25.7%	34.8%	13.6%
	50% Decrease	1	1	1	27.1%	19.9%	22.6%	14.9%
		1	1	2	6.1%	5.8%	9.3%	14.6%
		1	2	1	6.0%	2.9%	7.1%	13.1%
1		2	2	20.6%	35.8%	10.3%	15.0%	
1		2	3	6.7%	7.1%	7.7%	14.0%	
1		3	2	6.6%	4.5%	10.0%	14.7%	
1		3	1	26.6%	26.5%	33.0%	15.7%	
1/2-1/4-1/4 Information Across Stages	50% Increase	1	1	1	30.1%	24.6%	25.3%	14.6%
		1	1	2	3.2%	8.9%	3.6%	14.6%
		1	2	1	3.2%	0.4%	7.8%	14.0%
		1	2	2	26.6%	27.3%	16.0%	13.9%
		1	2	3	3.5%	11.1%	7.3%	14.0%
		1	3	2	3.5%	0.3%	4.0%	14.9%
		1	3	1	29.8%	26.6%	36.0%	15.0%
	Full	1	1	1	30.0%	25.4%	27.3%	13.6%
		1	1	2	3.3%	4.9%	5.6%	14.3%
		1	2	1	3.3%	2.0%	8.1%	14.1%
		1	2	2	26.4%	27.0%	12.3%	14.1%
		1	2	3	3.6%	4.4%	9.3%	14.2%
		1	3	2	3.6%	3.0%	4.0%	14.7%
		1	3	1	29.7%	33.4%	33.5%	15.0%
	25% Decrease	1	1	1	29.4%	19.6%	26.5%	13.7%
		1	1	2	4.0%	5.3%	7.7%	14.8%
		1	2	1	4.0%	1.6%	7.9%	14.1%
		1	2	2	25.4%	28.0%	12.3%	14.1%
		1	2	3	4.0%	5.6%	7.5%	14.3%
		1	3	2	4.0%	2.8%	5.8%	14.5%
		1	3	1	29.4%	36.5%	32.3%	14.4%
	50% Decrease	1	1	1	28.8%	24.4%	21.6%	14.8%
		1	1	2	4.5%	3.5%	8.9%	14.7%
		1	2	1	4.5%	1.6%	8.8%	14.5%
1		2	2	24.3%	25.4%	9.2%	14.0%	
1		2	3	4.5%	5.7%	10.8%	13.9%	
1		3	2	4.5%	3.3%	8.1%	15.0%	
1		3	1	28.8%	33.8%	32.5%	14.2%	

Table 4.25. Routing Path Frequencies in 1-2-3 Design with Four Routing Strategies

Division of Information	TIF Level	Module			Routing Strategy			
		s1	s2	s3	DPI	Proximity	NC	Random
Equal Information Across Stages	50% Increase	1	1	1	33.0%	33.3%	35.2%	16.9%
		1	1	2	16.7%	16.7%	12.1%	16.7%
		1	1	3	0.2%	0.0%	1.5%	16.6%
		1	2	1	0.3%	0.1%	0.1%	16.6%
		1	2	2	16.6%	16.6%	11.0%	16.4%
		1	2	3	33.1%	33.3%	40.1%	16.8%
	Full	1	1	1	32.7%	33.3%	35.1%	16.9%
		1	1	2	16.7%	16.7%	10.6%	16.9%
		1	1	3	0.5%	0.1%	0.3%	16.6%
		1	2	1	0.6%	0.1%	0.4%	16.7%
		1	2	2	16.6%	16.7%	12.4%	16.3%
		1	2	3	32.8%	33.3%	41.2%	16.7%
	25% Decrease	1	1	1	32.3%	33.1%	26.6%	16.8%
		1	1	2	16.8%	16.7%	8.4%	16.6%
		1	1	3	0.9%	0.2%	6.5%	16.5%
		1	2	1	1.0%	0.2%	0.8%	16.4%
		1	2	2	16.5%	16.6%	15.3%	16.8%
		1	2	3	32.5%	33.2%	42.3%	16.9%
	50% Decrease	1	1	1	31.9%	32.9%	35.6%	16.4%
		1	1	2	16.5%	16.8%	10.0%	16.7%
		1	1	3	1.6%	0.4%	1.1%	16.7%
		1	2	1	1.5%	0.5%	1.7%	16.6%
		1	2	2	16.8%	16.6%	13.2%	16.7%
		1	2	3	31.8%	33.0%	40.4%	16.9%
1/2-1/4-1/4 Information Across Stages	50% Increase	1	1	1	33.3%	30.2%	25.3%	16.7%
		1	1	2	16.7%	12.1%	15.1%	16.6%
		1	1	3	0.0%	1.5%	4.4%	16.8%
		1	2	1	0.1%	0.1%	4.4%	16.8%
		1	2	2	16.6%	16.0%	12.5%	16.7%
		1	2	3	33.3%	40.1%	38.3%	16.5%
	Full	1	1	1	33.3%	28.1%	25.7%	16.9%
		1	1	2	16.7%	15.6%	13.9%	16.5%
		1	1	3	0.1%	0.3%	7.6%	16.6%
		1	2	1	0.1%	0.4%	4.9%	16.7%
		1	2	2	16.7%	17.4%	14.0%	16.6%
		1	2	3	33.3%	38.2%	34.1%	16.7%
	25% Decrease	1	1	1	33.1%	29.6%	23.7%	17.0%
		1	1	2	16.7%	18.4%	11.6%	16.7%
		1	1	3	0.2%	1.5%	6.8%	16.6%
		1	2	1	0.2%	0.8%	8.2%	16.6%
		1	2	2	16.6%	15.3%	13.7%	16.6%
		1	2	3	33.2%	34.3%	36.0%	16.6%
	50% Decrease	1	1	1	32.9%	35.6%	21.3%	16.6%
		1	1	2	16.8%	11.0%	11.9%	16.5%
		1	1	3	0.4%	1.1%	9.8%	16.5%
		1	2	1	0.5%	1.7%	8.0%	16.7%
		1	2	2	16.6%	13.2%	12.6%	16.9%
		1	2	3	33.0%	37.4%	36.4%	16.8%

Table 4.26. Routing Path Frequencies in 1-3-2 Design with Four Routing Strategies

Division of Information	TIF Level	Module			Routing Strategy			
		s1	s2	s3	DPI	Proximity	NC	Random
Equal Information Across Stages	50% Increase	1	1	1	33.0%	33.3%	31.6%	16.8%
		1	1	2	0.3%	0.1%	2.7%	16.5%
		1	2	1	16.6%	16.6%	13.7%	16.6%
		1	2	2	16.7%	16.7%	15.1%	16.8%
		1	3	1	0.4%	0.1%	0.0%	16.8%
		1	3	2	33.0%	33.2%	36.8%	16.5%
	Full	1	1	1	32.8%	33.2%	29.1%	16.6%
		1	1	2	0.5%	0.1%	1.2%	16.9%
		1	2	1	16.6%	16.7%	17.3%	16.5%
		1	2	2	16.7%	16.7%	14.0%	16.7%
		1	3	1	0.6%	0.1%	0.2%	16.8%
		1	3	2	32.7%	33.2%	38.2%	16.5%
	25% Decrease	1	1	1	32.4%	33.1%	22.8%	16.4%
		1	1	2	1.0%	0.2%	2.6%	16.9%
		1	2	1	16.6%	16.6%	14.9%	16.8%
		1	2	2	16.8%	16.7%	10.4%	16.7%
		1	3	1	1.1%	0.3%	0.4%	16.6%
		1	3	2	32.3%	33.1%	49.0%	16.6%
	50% Decrease	1	1	1	31.9%	32.9%	16.7%	16.6%
		1	1	2	1.5%	0.4%	3.3%	16.5%
		1	2	1	16.4%	16.5%	13.3%	16.5%
		1	2	2	17.0%	16.8%	19.3%	16.8%
		1	3	1	1.8%	0.5%	0.6%	16.7%
		1	3	2	31.6%	32.8%	46.8%	16.9%
1/2-1/4-1/4 Information Across Stages	50% Increase	1	1	1	33.3%	31.6%	25.3%	16.9%
		1	1	2	0.1%	2.7%	3.6%	16.4%
		1	2	1	16.6%	8.7%	15.9%	16.5%
		1	2	2	16.7%	15.1%	15.2%	16.9%
		1	3	1	0.1%	0.0%	4.0%	16.8%
		1	3	2	33.2%	41.8%	36.0%	16.6%
	Full	1	1	1	33.2%	29.1%	27.3%	16.6%
		1	1	2	0.1%	1.2%	5.6%	17.0%
		1	2	1	16.7%	17.3%	11.9%	16.5%
		1	2	2	16.7%	14.0%	17.8%	16.6%
		1	3	1	0.1%	0.2%	4.0%	16.5%
		1	3	2	33.2%	38.2%	33.5%	16.8%
	25% Decrease	1	1	1	33.1%	32.8%	26.5%	16.6%
		1	1	2	0.2%	2.6%	7.7%	16.8%
		1	2	1	16.6%	14.9%	11.4%	16.6%
		1	2	2	16.7%	10.4%	13.3%	16.8%
		1	3	1	0.3%	0.4%	5.8%	16.6%
		1	3	2	33.1%	39.0%	35.3%	16.6%
	50% Decrease	1	1	1	32.9%	26.7%	21.6%	16.5%
		1	1	2	0.4%	3.3%	8.9%	16.7%
		1	2	1	16.5%	13.3%	13.2%	16.6%
		1	2	2	16.8%	19.3%	15.6%	16.8%
		1	3	1	0.5%	0.6%	8.1%	16.8%
		1	3	2	32.8%	36.8%	32.5%	16.6%

Figure 4.1. RMSEs for DPI Routing with 1-2-2 Design at Three Pass Rates

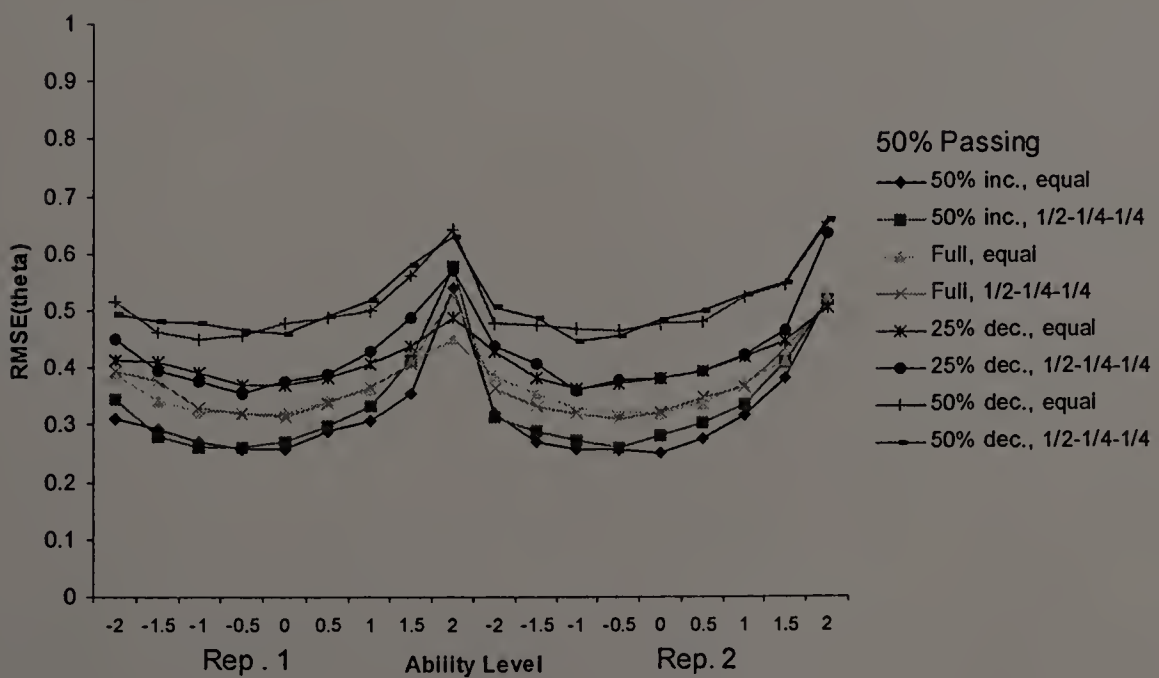
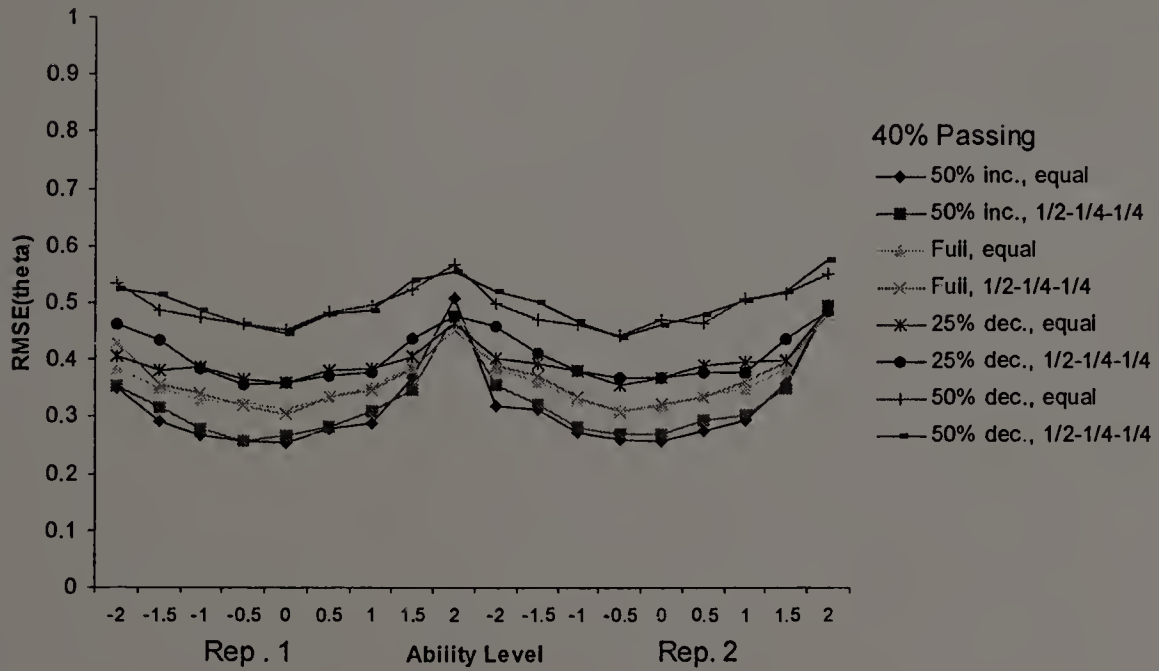
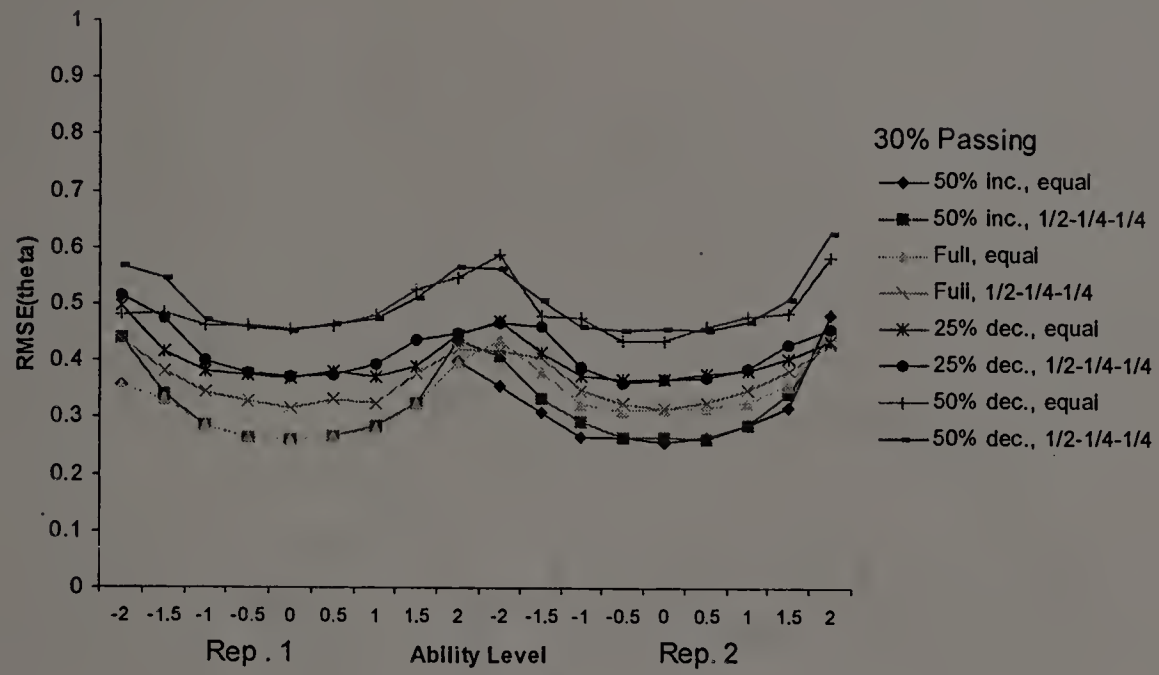


Figure 4.2. RMSEs for DPI Routing with 1-3-3 Design at Three Pass Rates

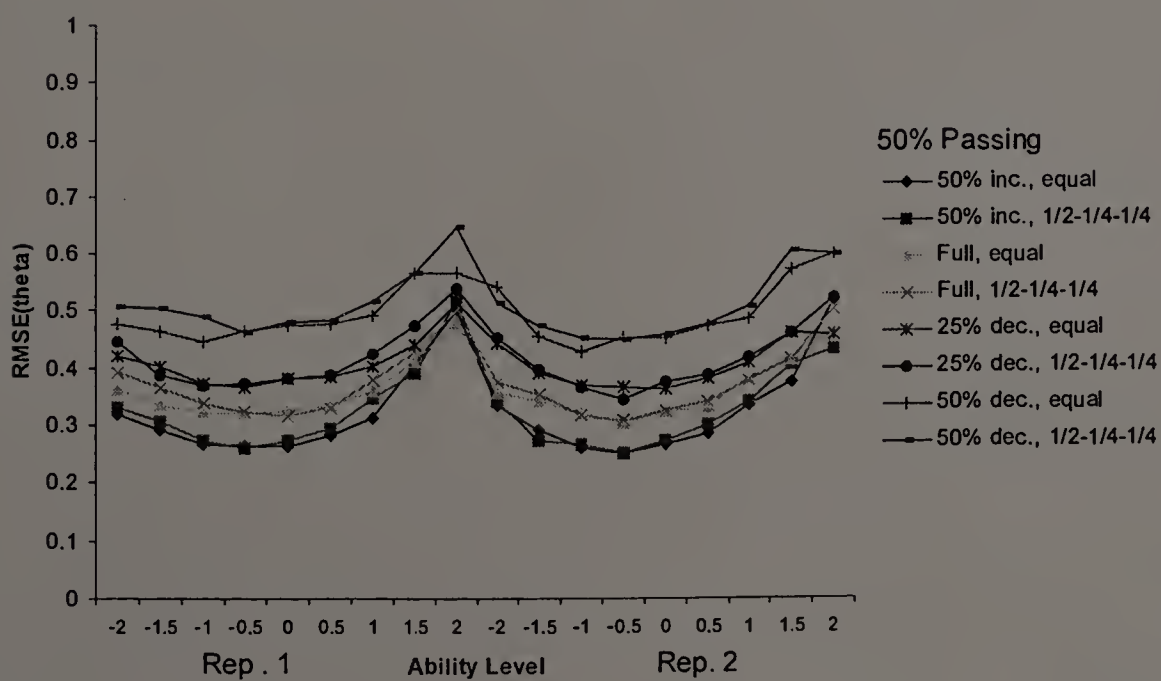
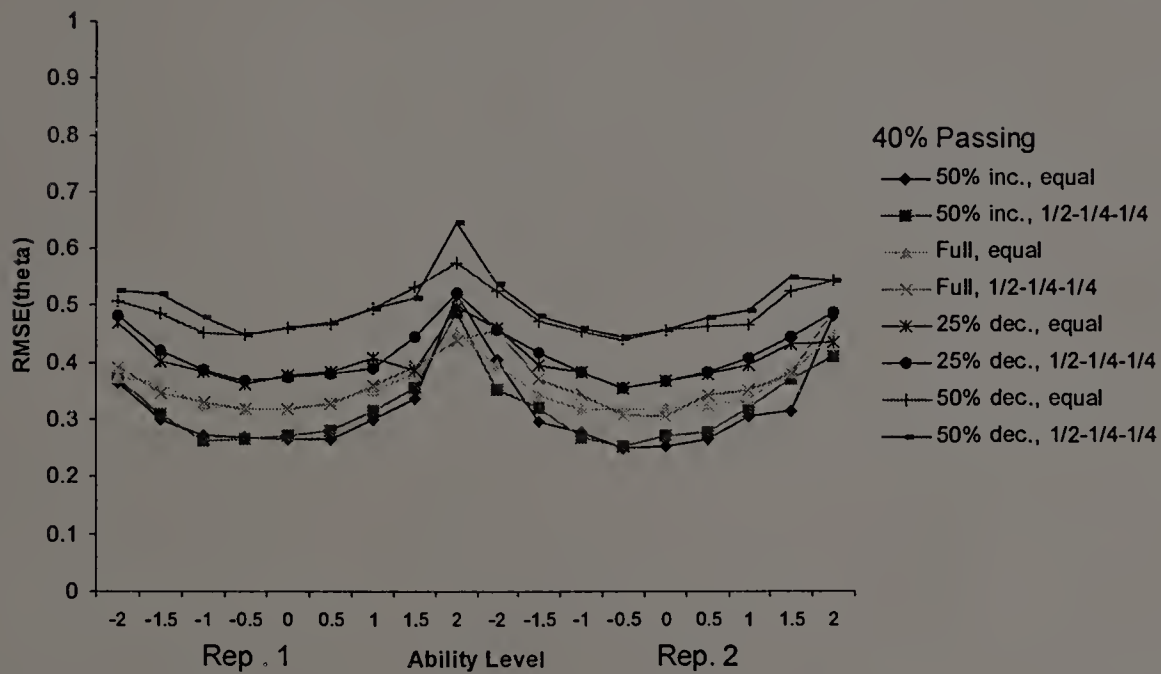
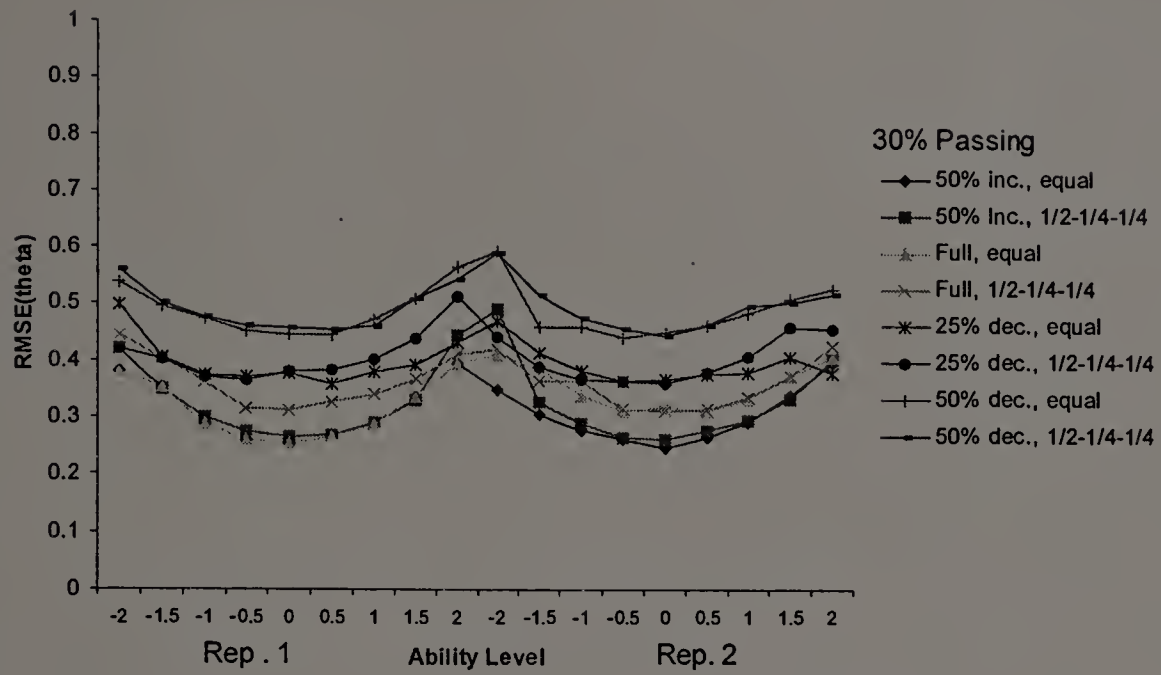


Figure 4.3. RMSEs for DPI Routing with 1-2-3 Design at Three Pass Rates

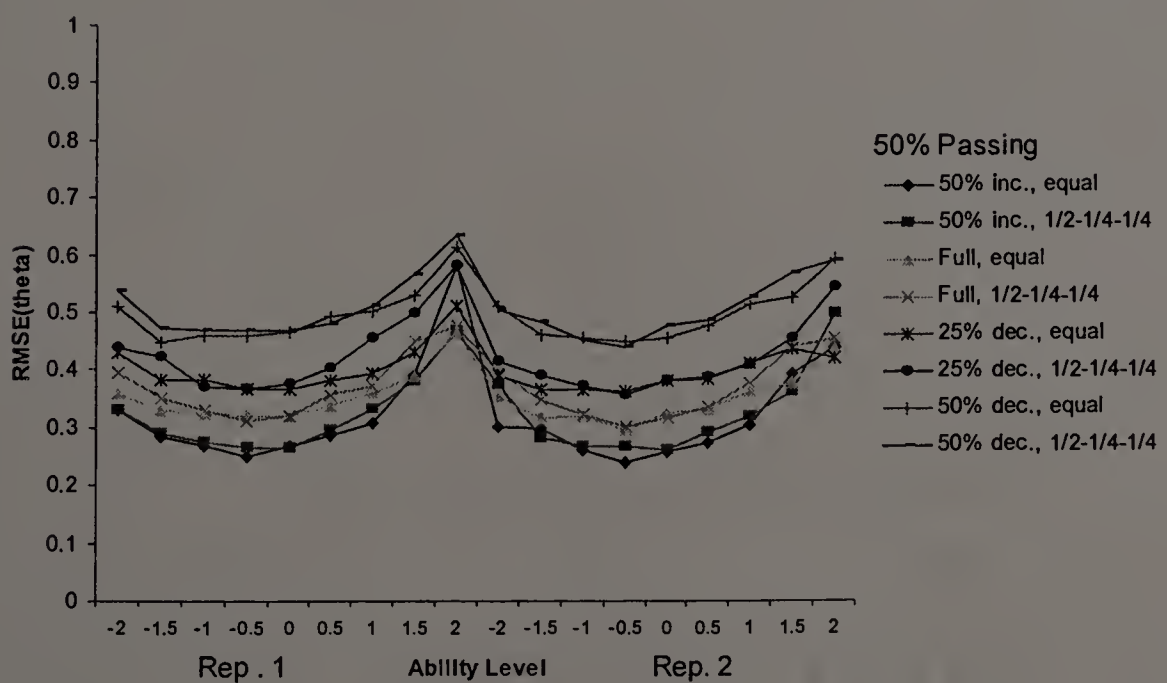
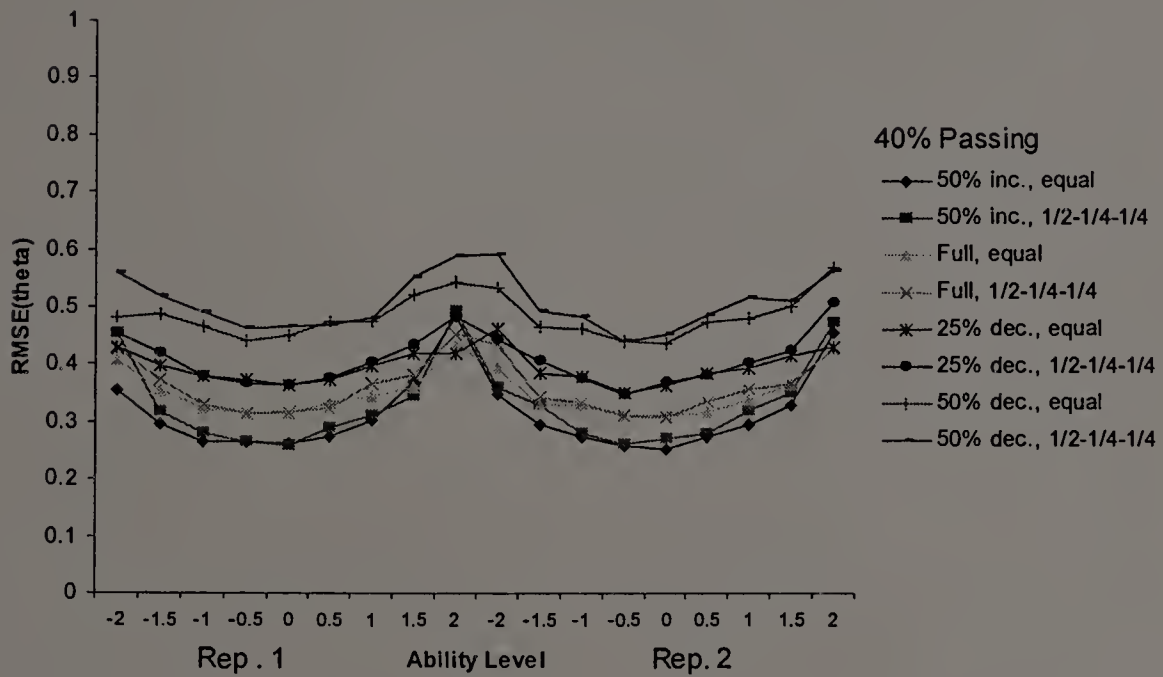
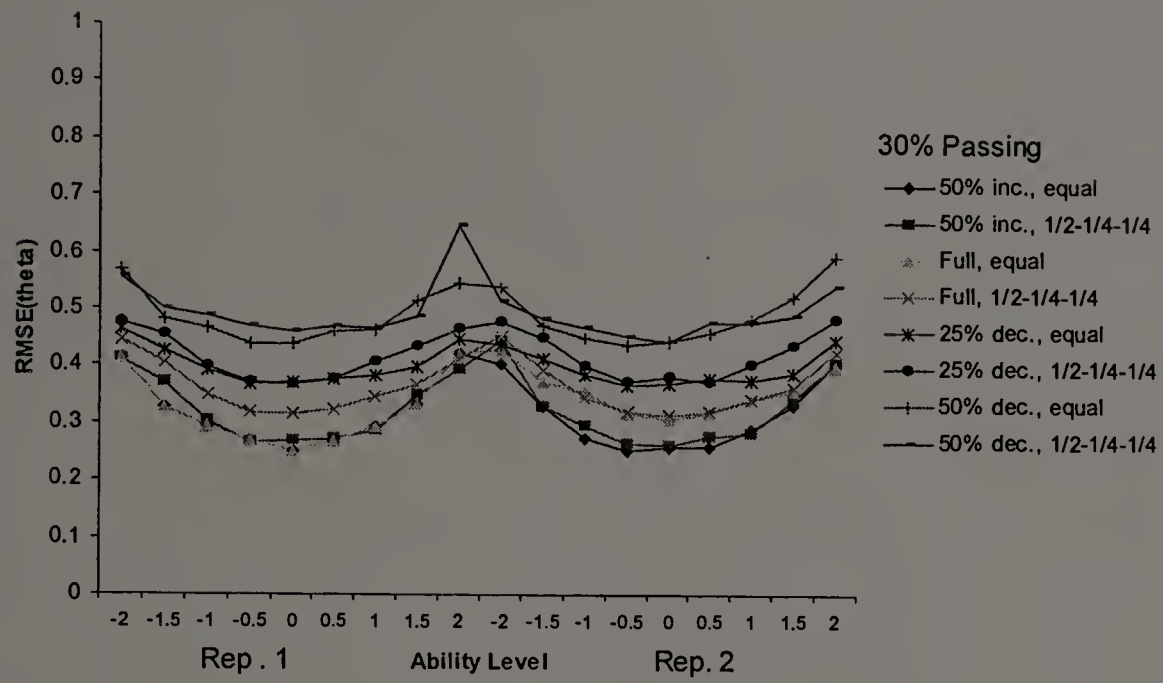


Figure 4.4. RMSEs for DPI Routing with 1-3-2 Design at Three Pass Rates

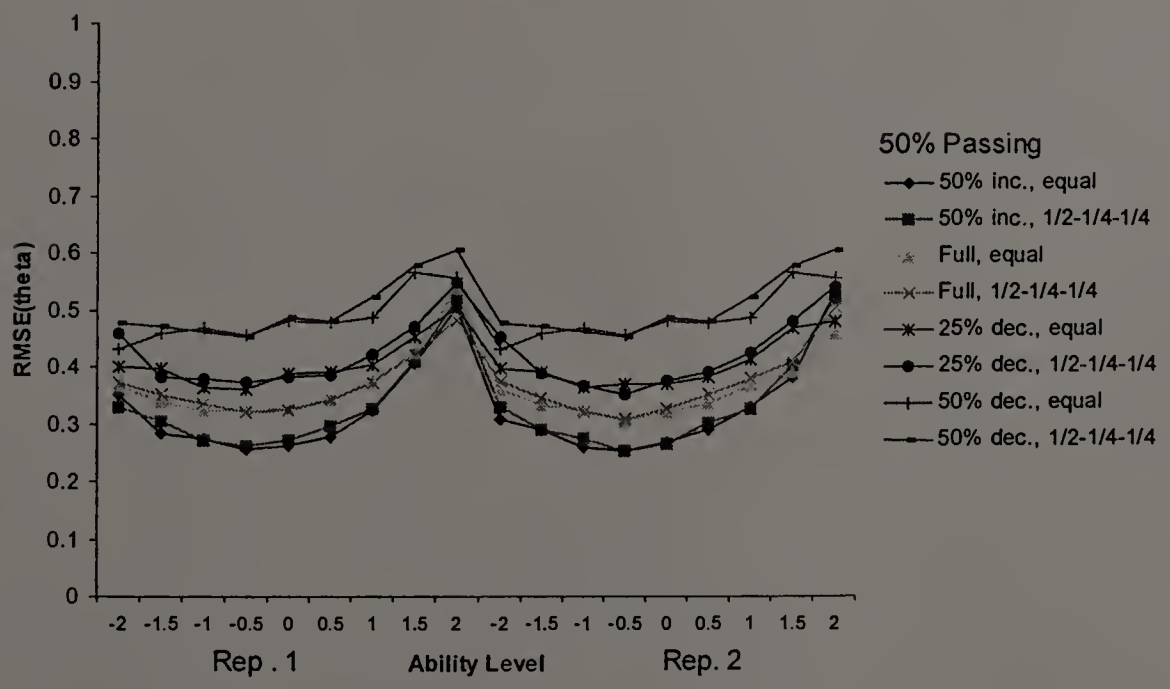
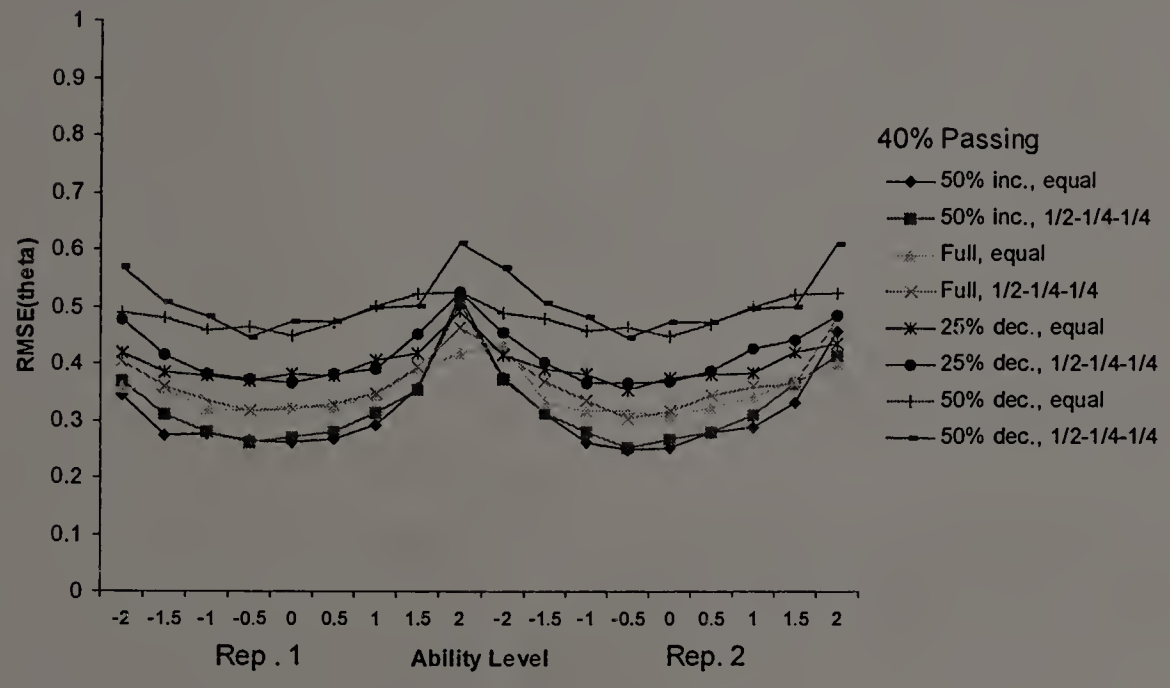
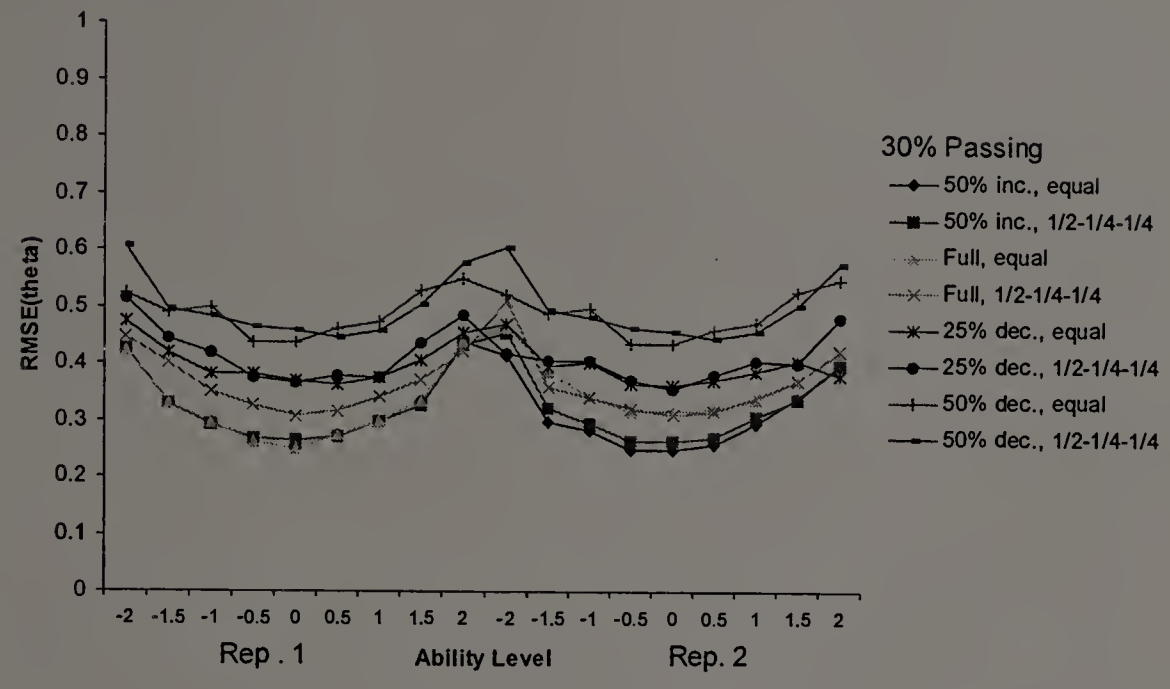


Figure 4.5. RMSEs for Proximity Routing with 1-2-2 Design at Three Pass Rates

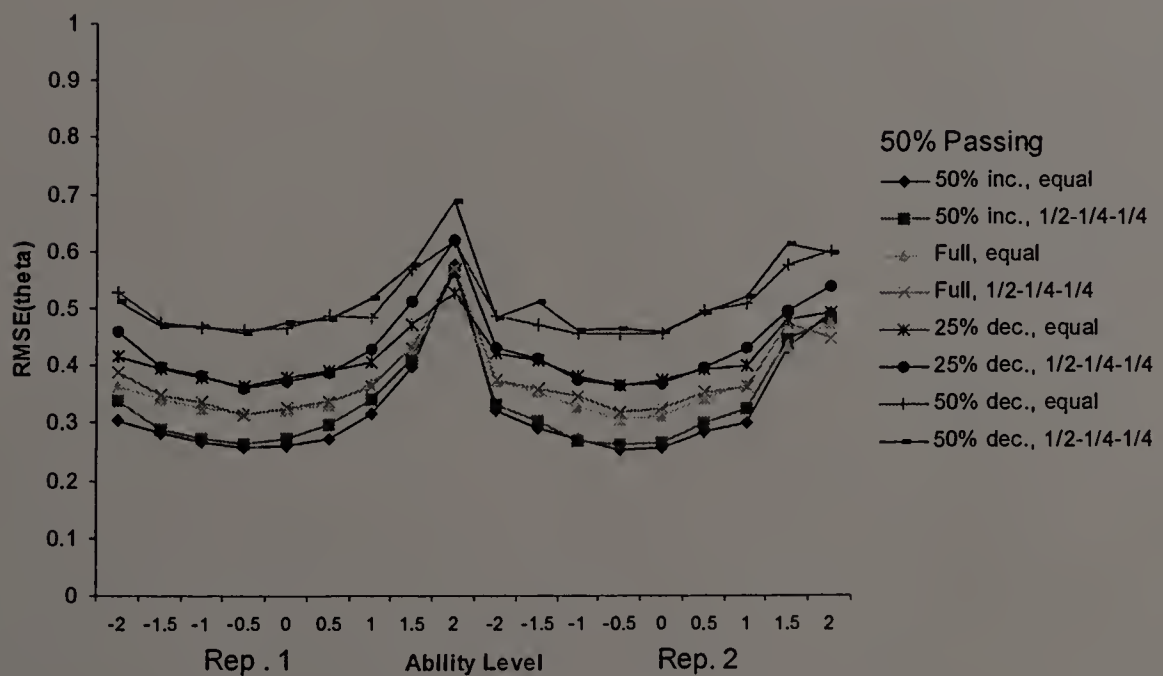
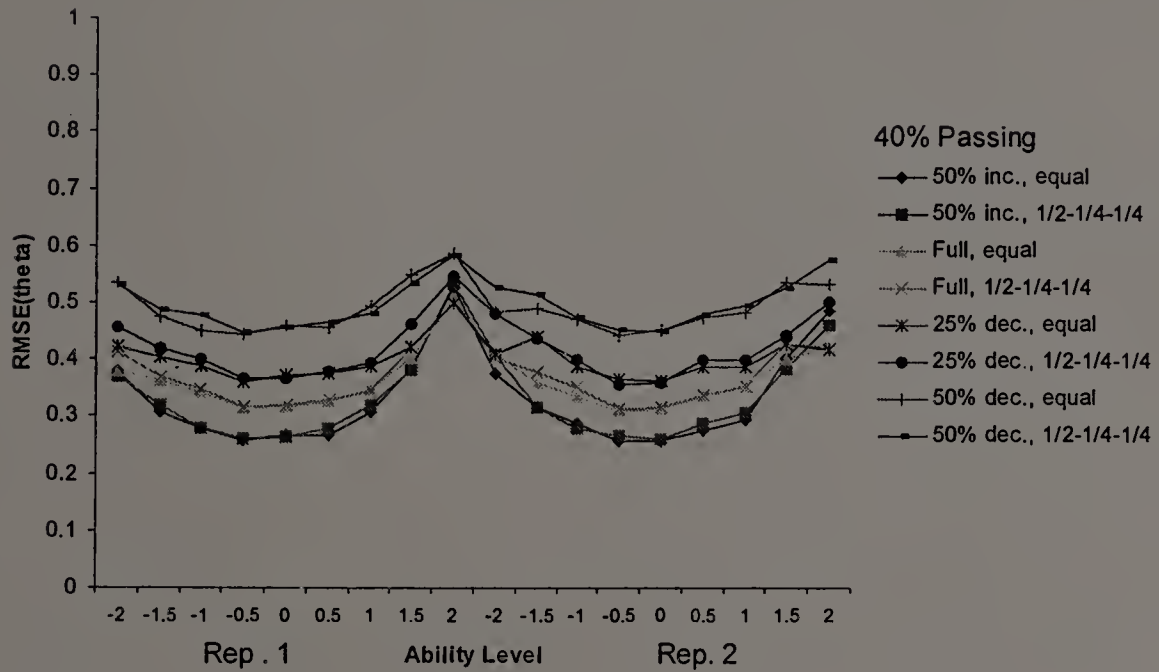
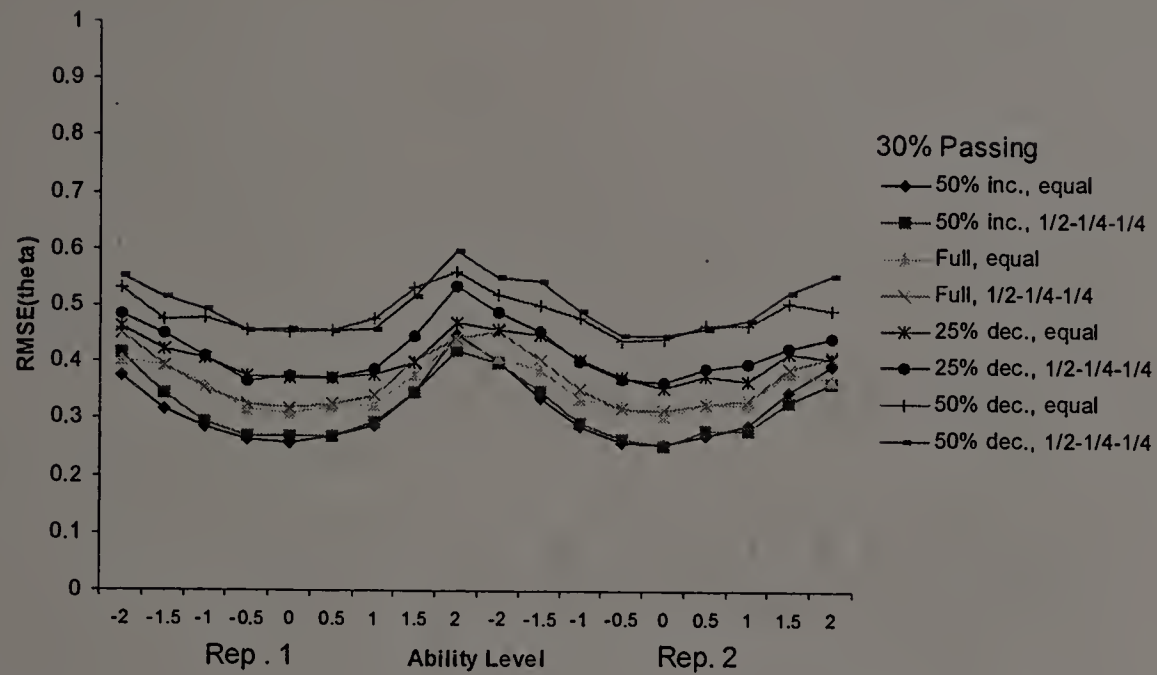


Figure 4.6. RMSEs for Proximity Routing with 1-3-3 Design at Three Pass Rates

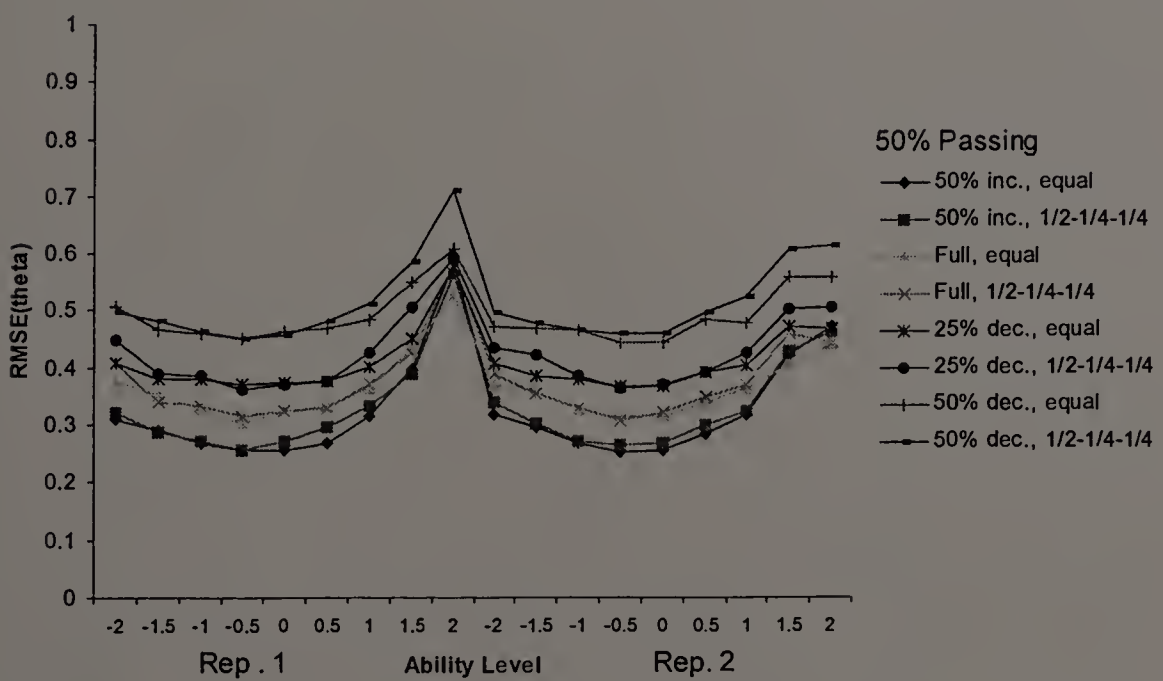
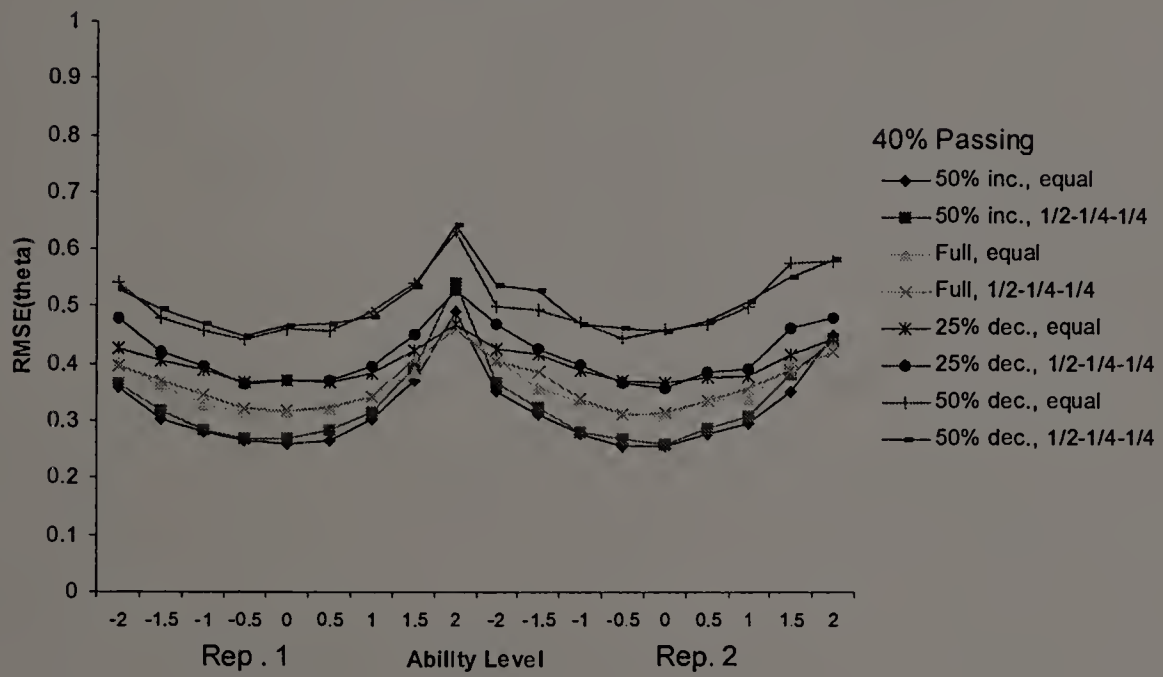
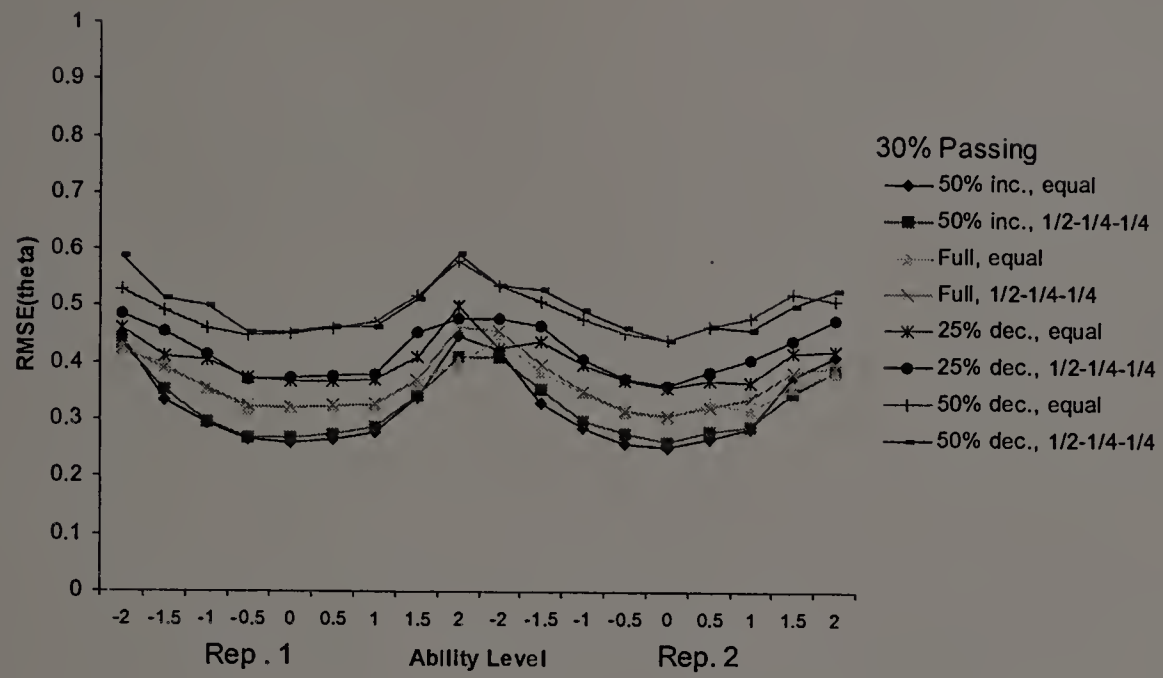


Figure 4.7. RMSEs for Proximity Routing with 1-2-3 Design at Three Pass Rates

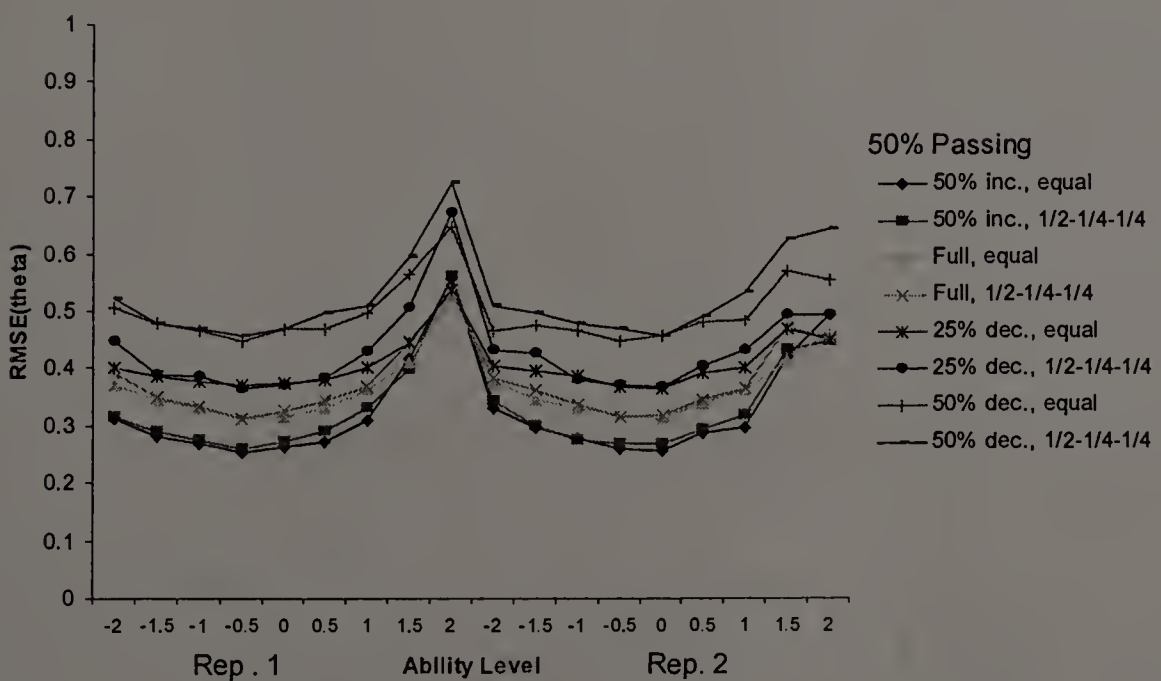
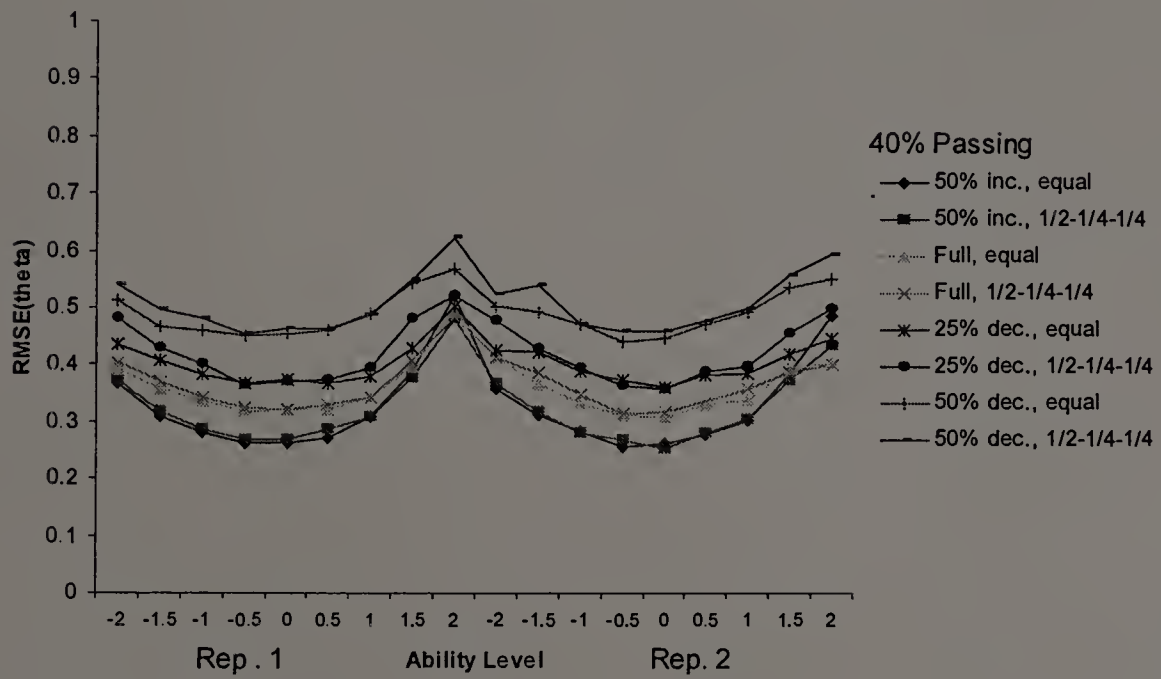
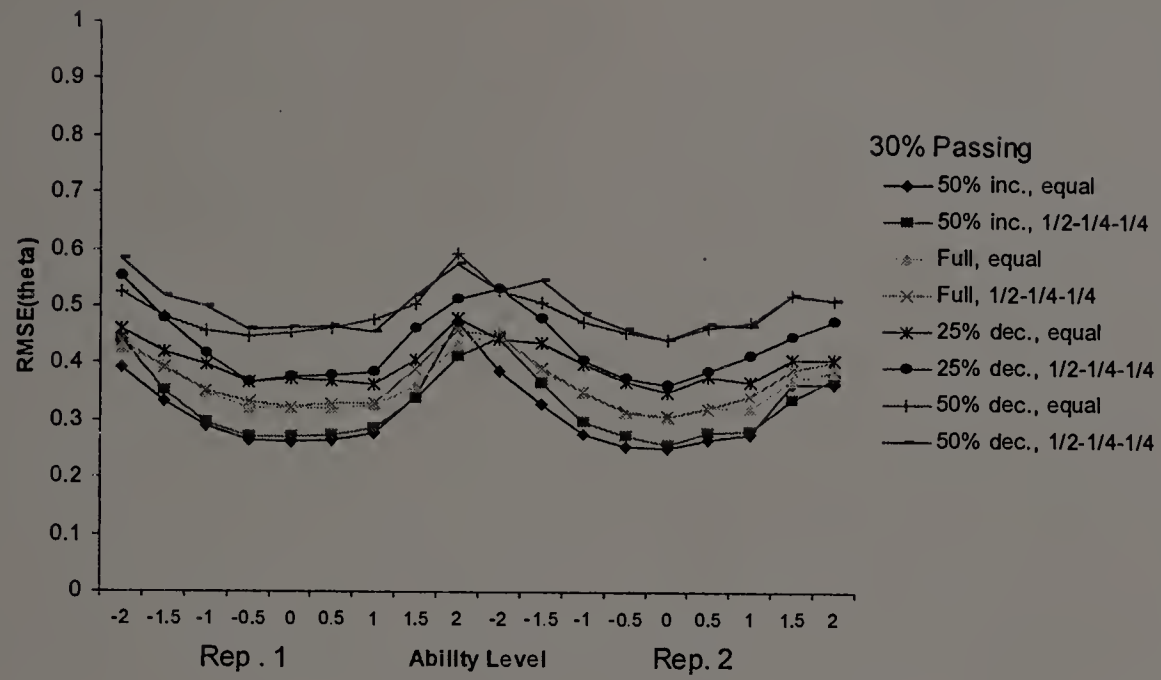


Figure 4.8. RMSEs for Proximity Routing with 1-3-2 Design at Three Pass Rates

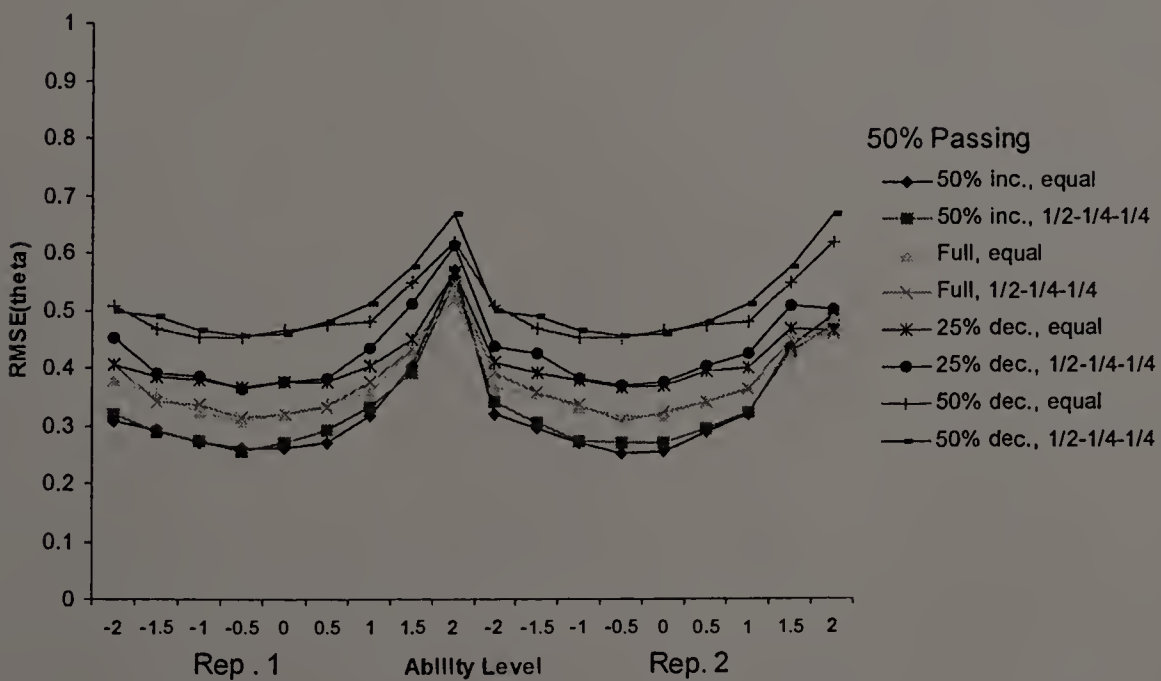
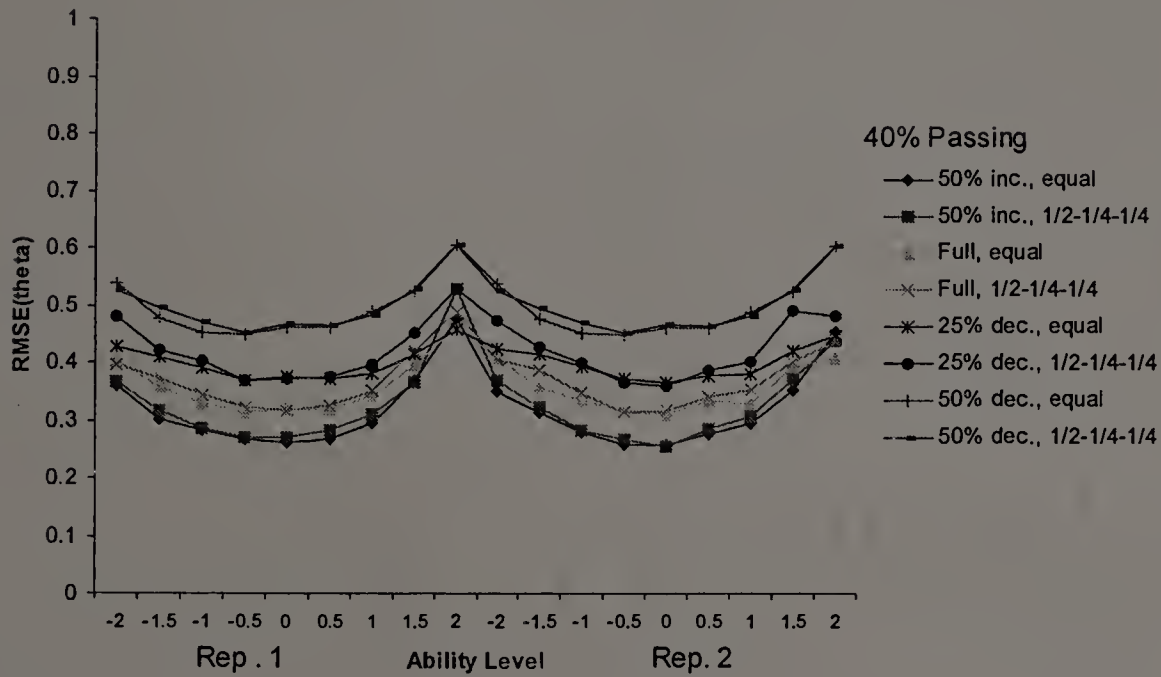
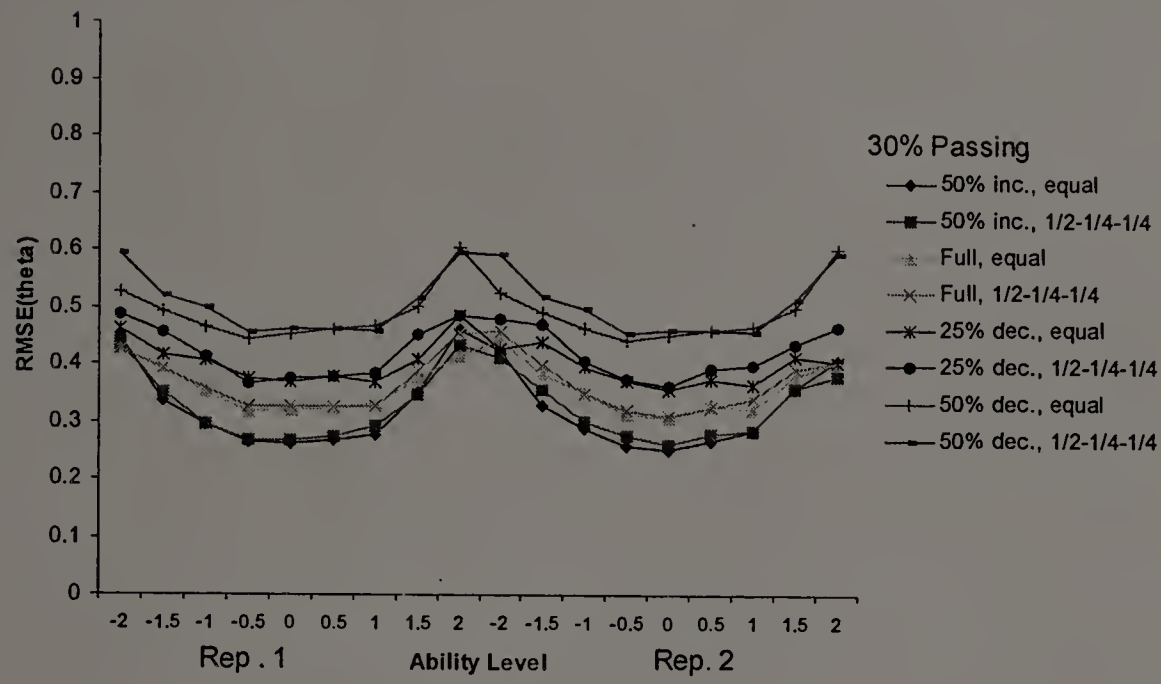


Figure 4.9. RMSEs for NC Routing with 1-2-2 Design at Three Pass Rates

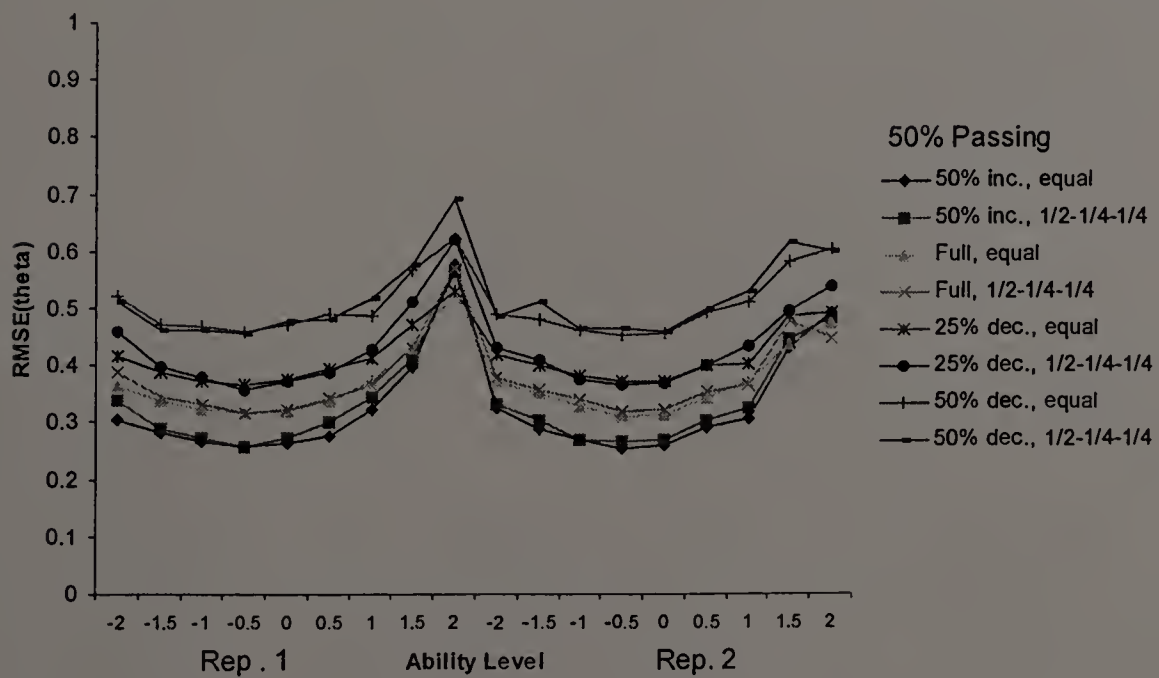
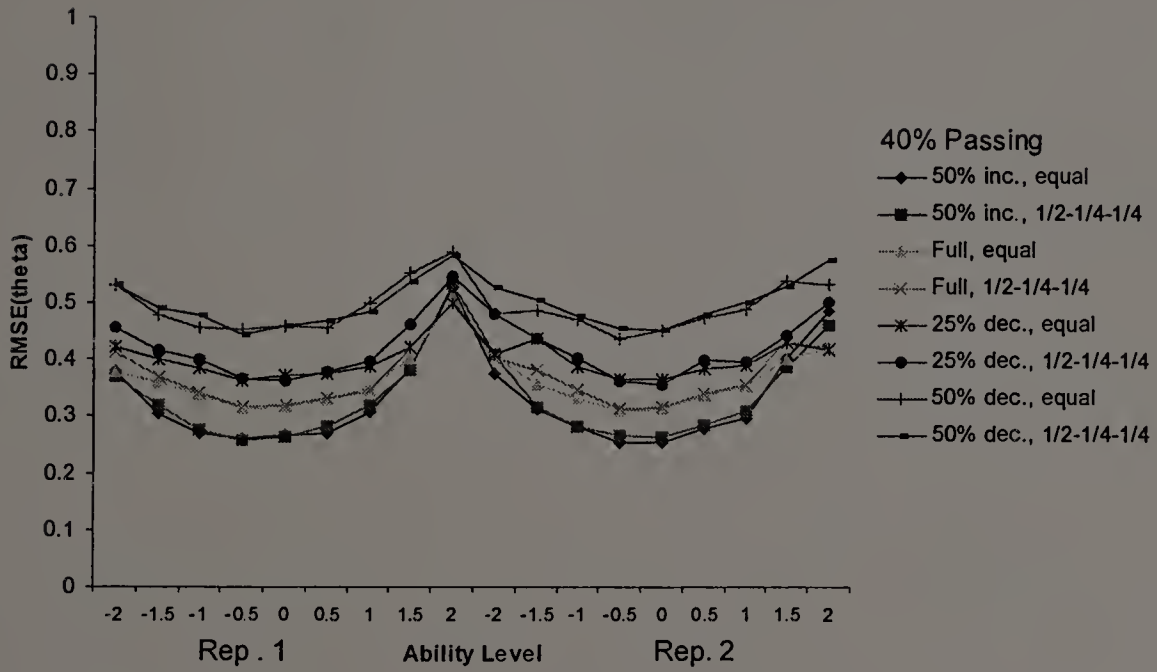
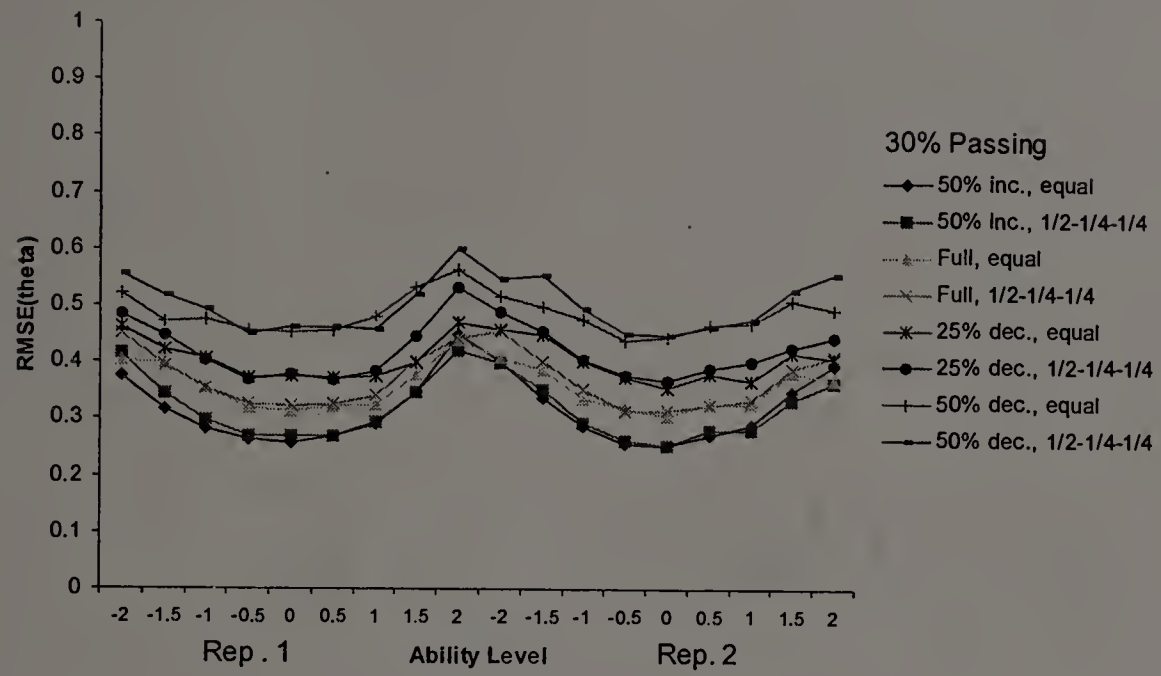


Figure 4.10. RMSEs for NC Routing with 1-3-3 Design at Three Pass Rates

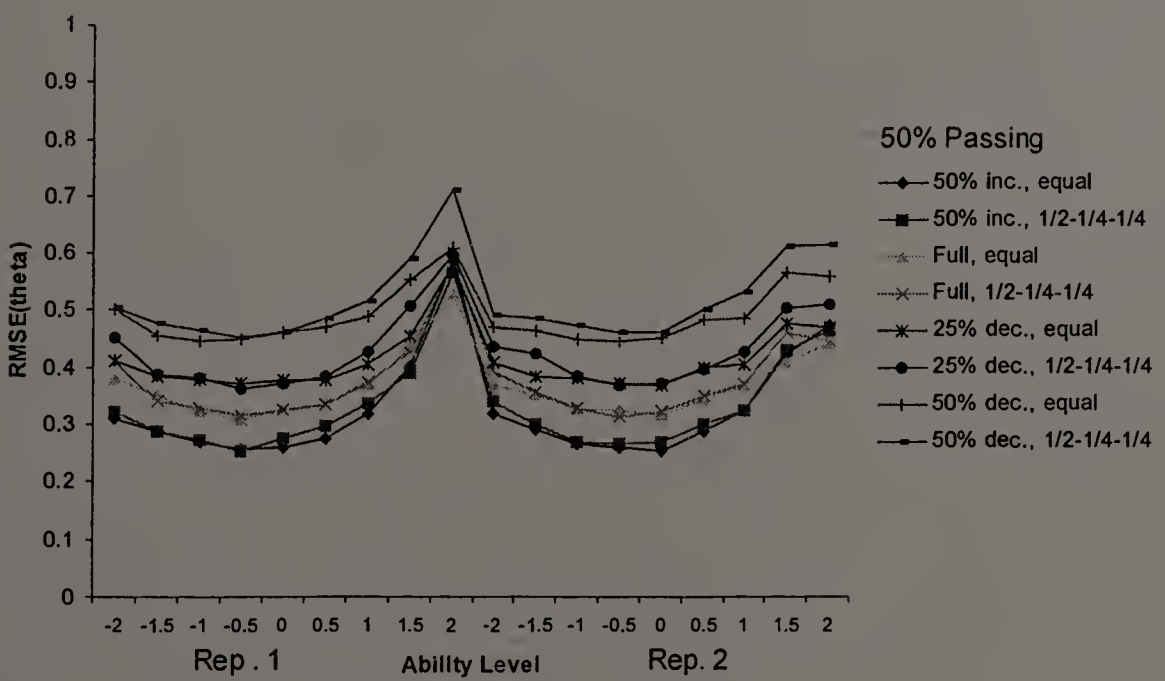
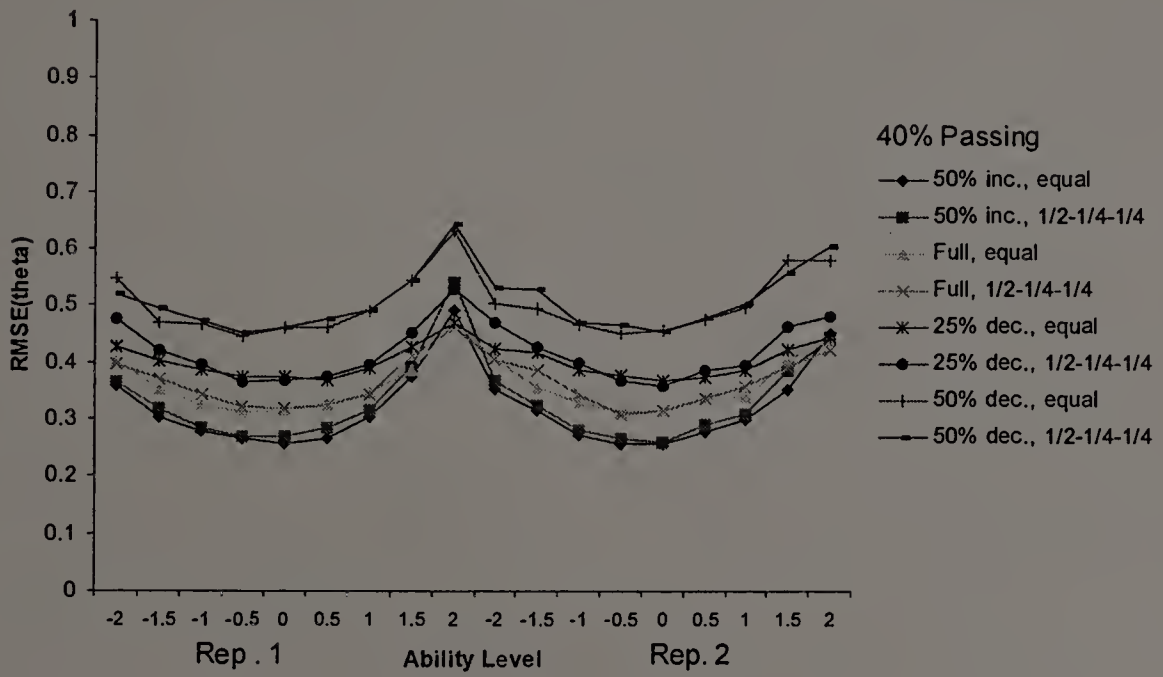
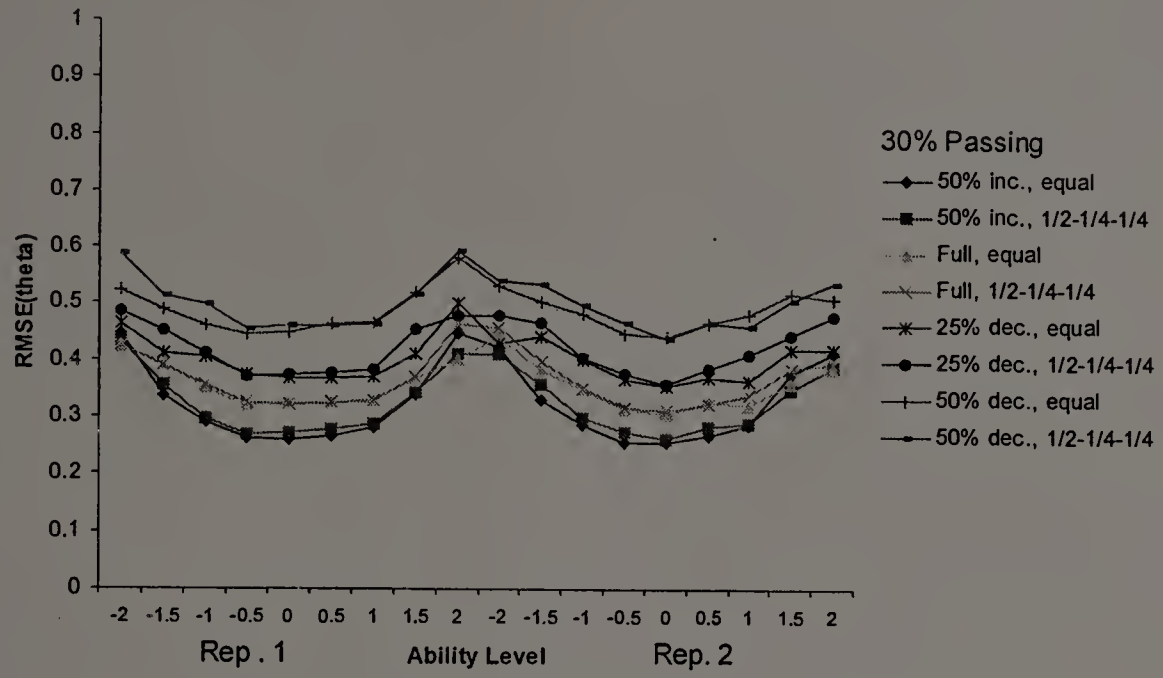


Figure 4.11. RMSEs for NC Routing with 1-2-3 Design at Three Pass Rates

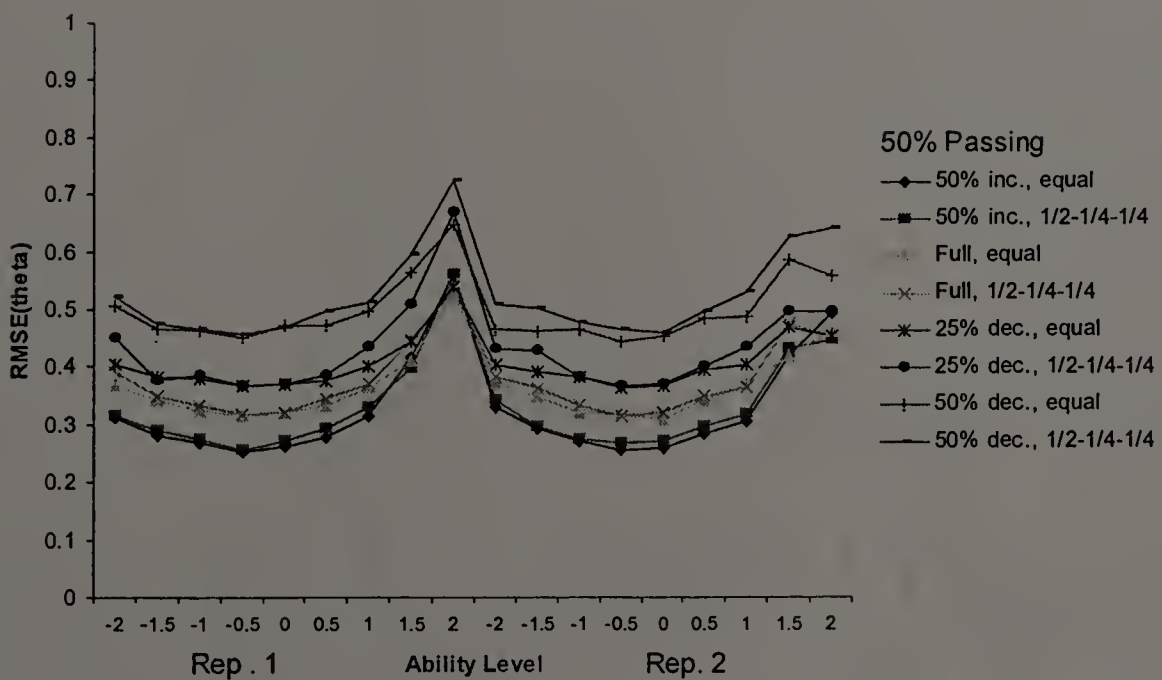
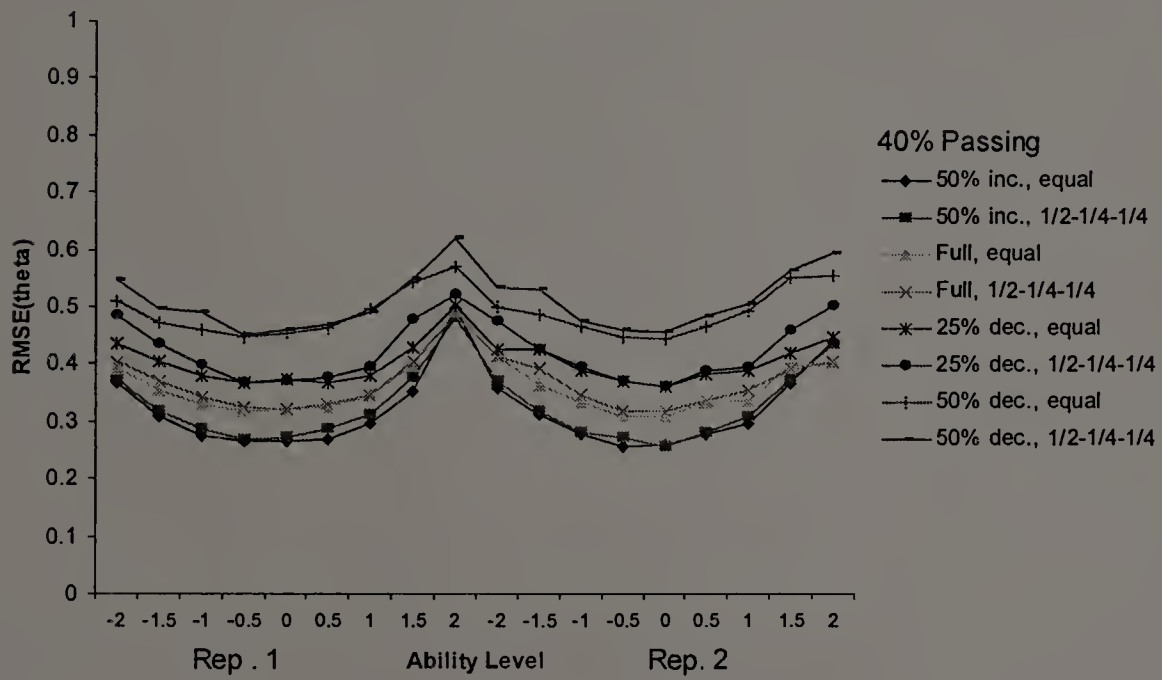
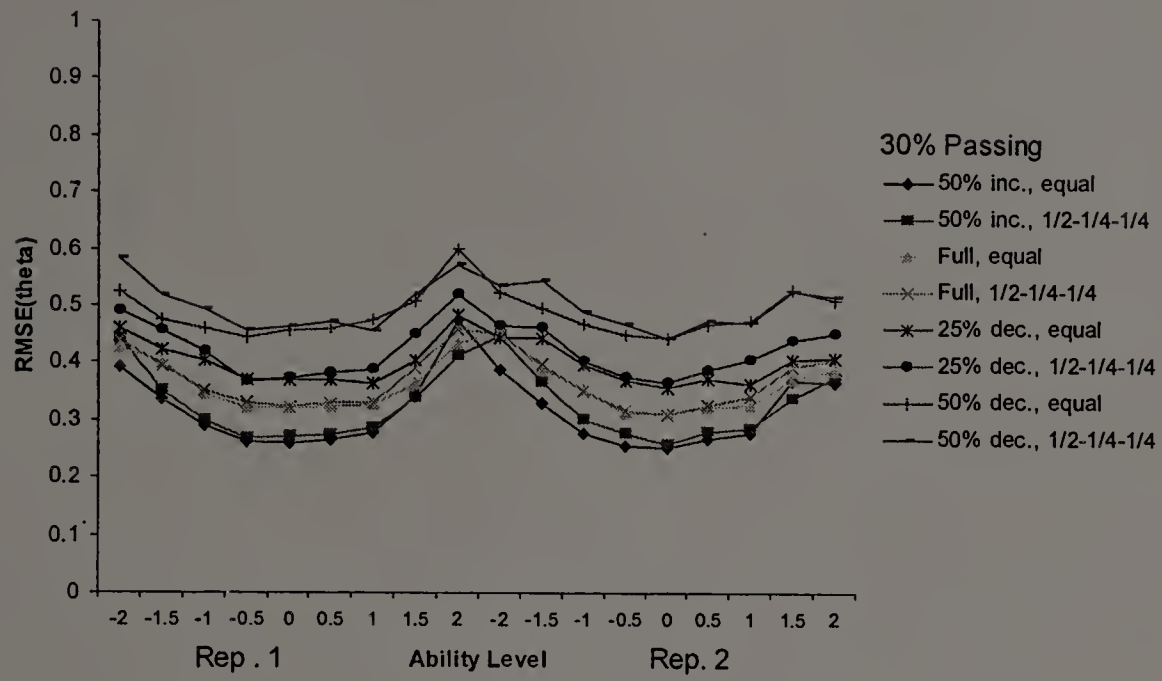


Figure 4.12. RMSEs for NC Routing with 1-3-2 Design at Three Pass Rates

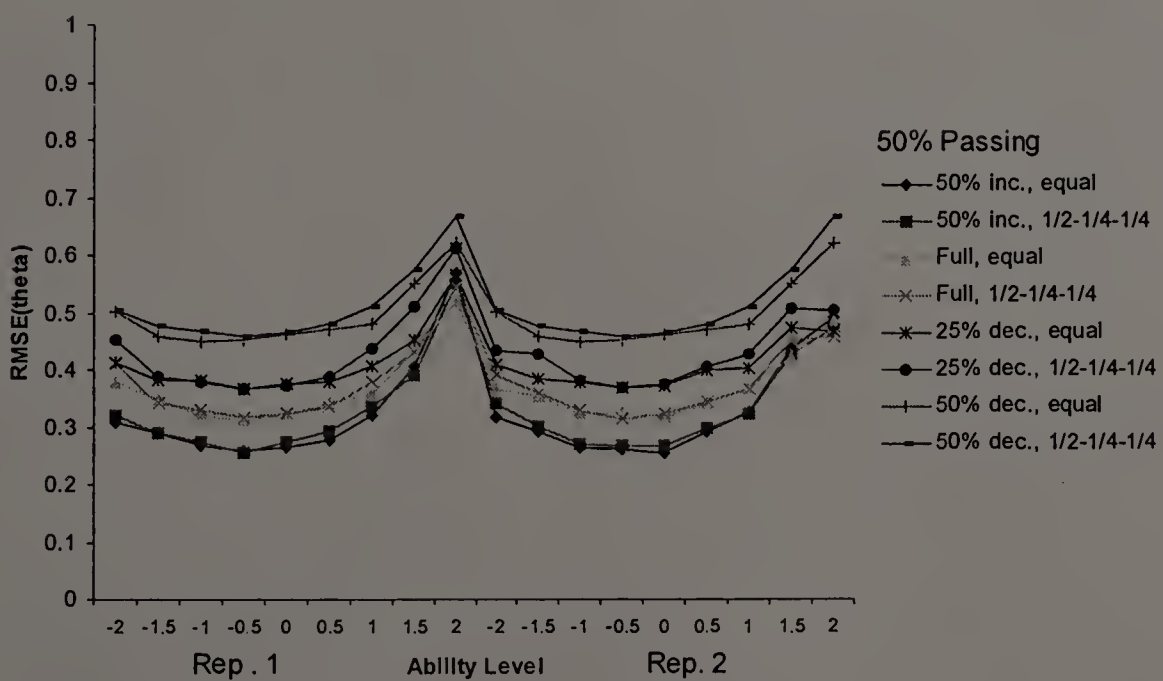
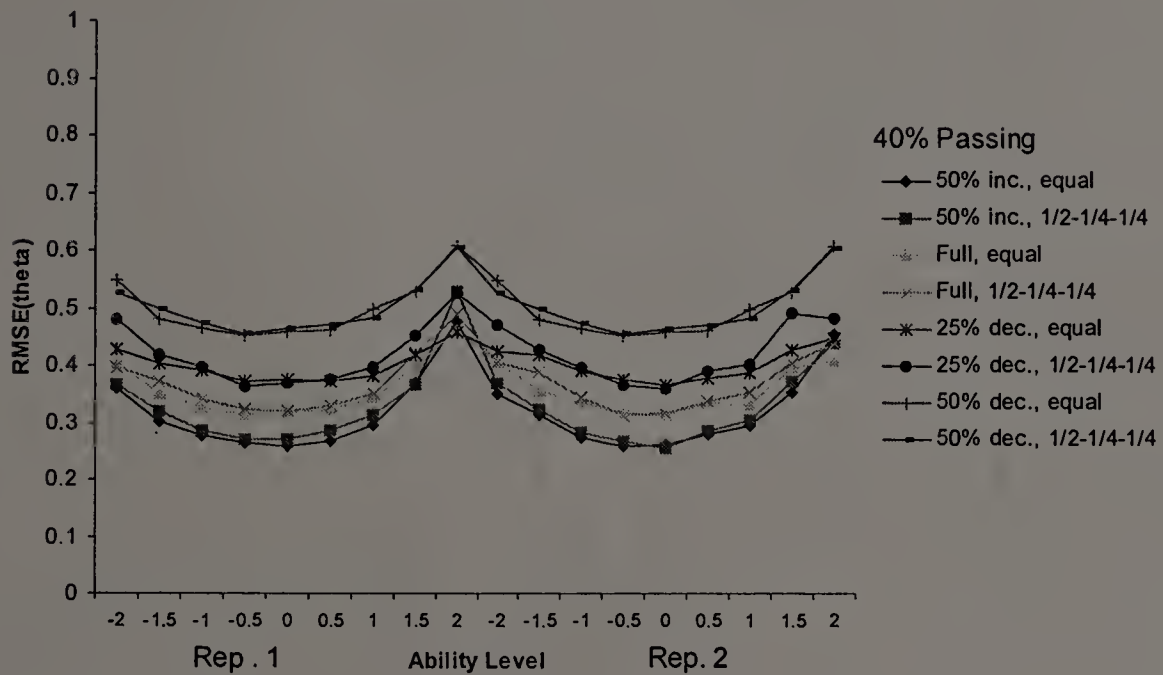
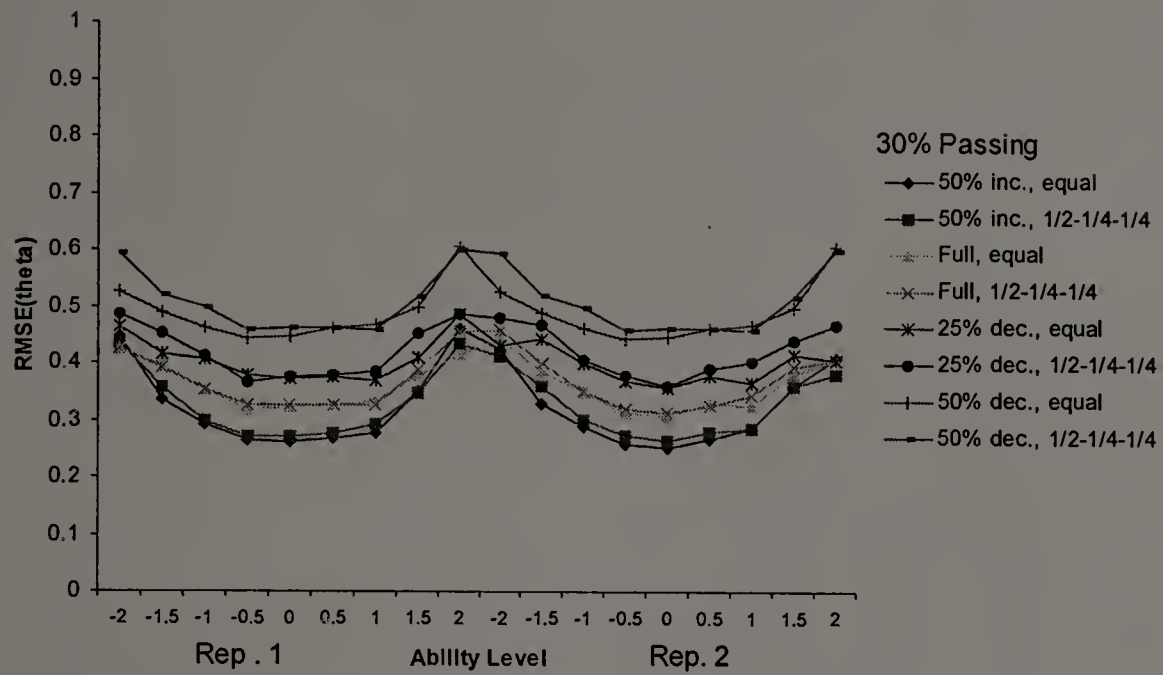


Figure 4.13. RMSEs for Random Routing with 1-2-2 Design at Three Pass Rates

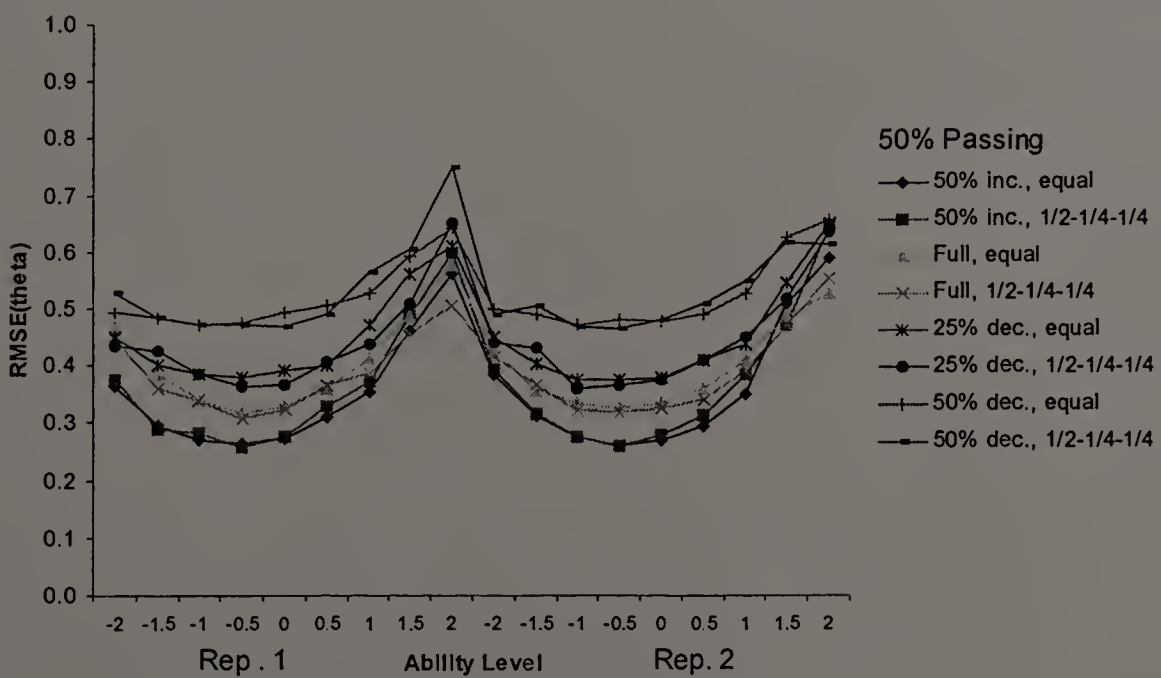
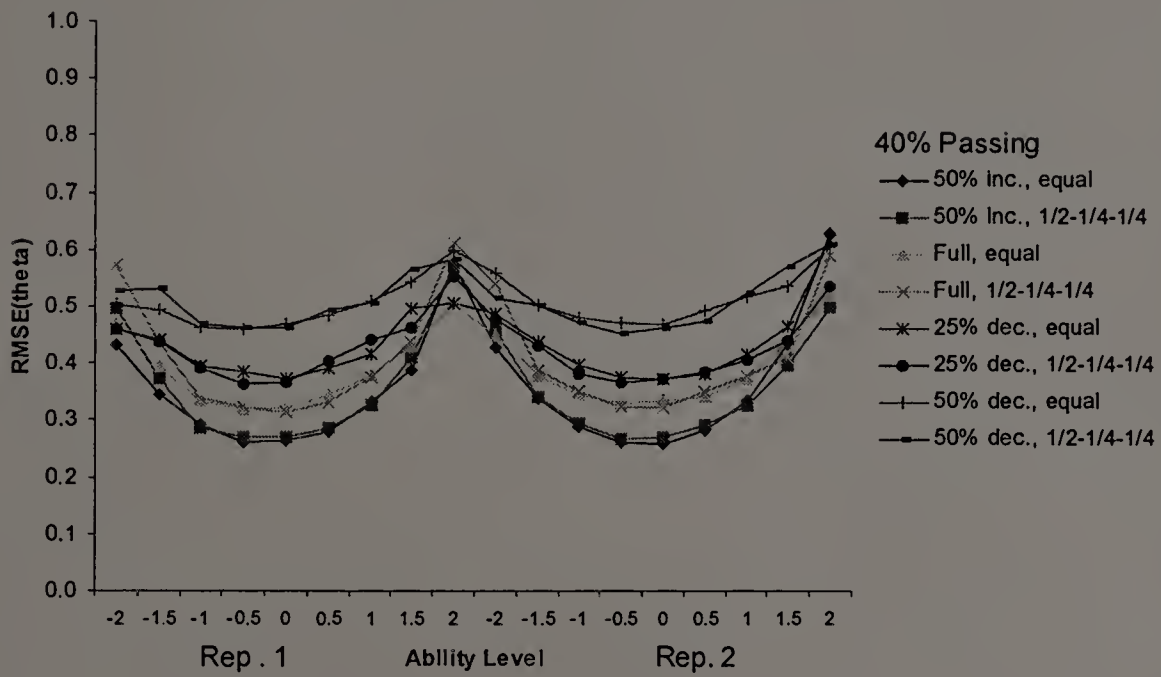
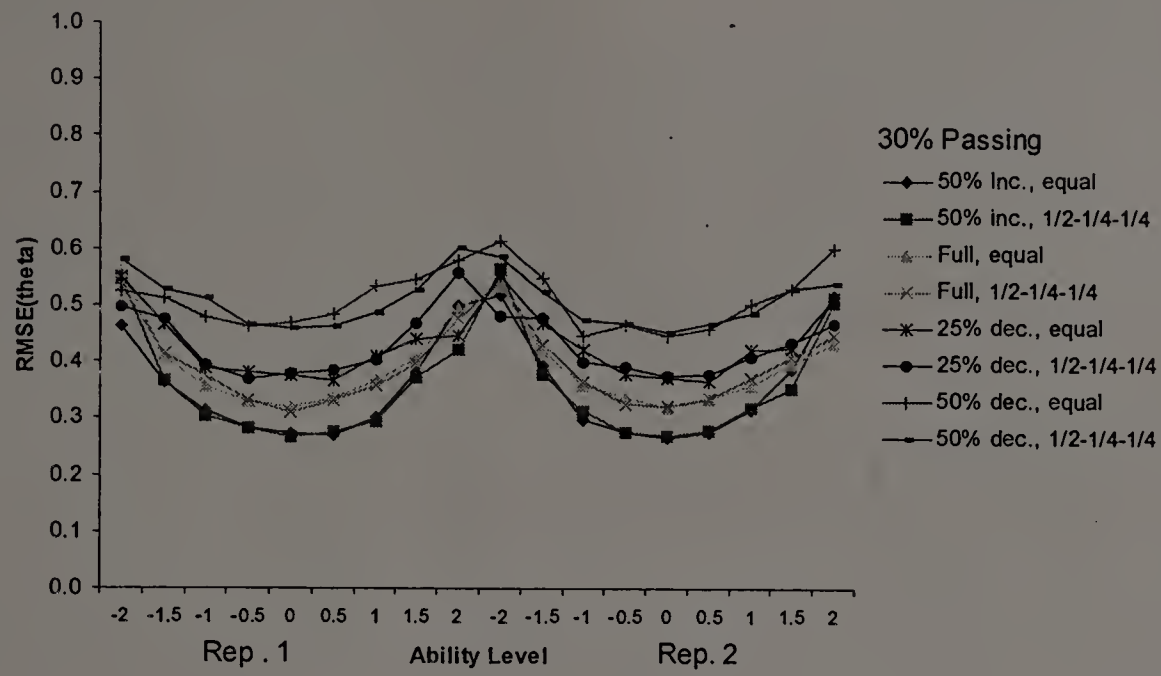


Figure 4.14. RMSEs for Random Routing with 1-3-3 Design at Three Pass Rates

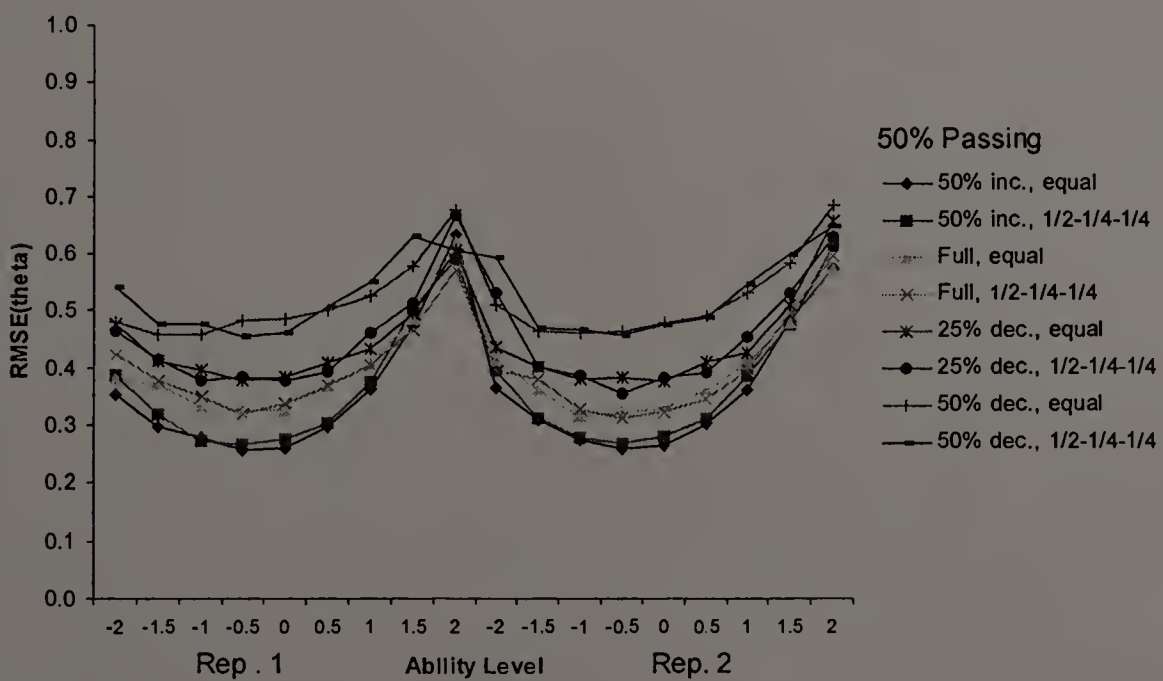
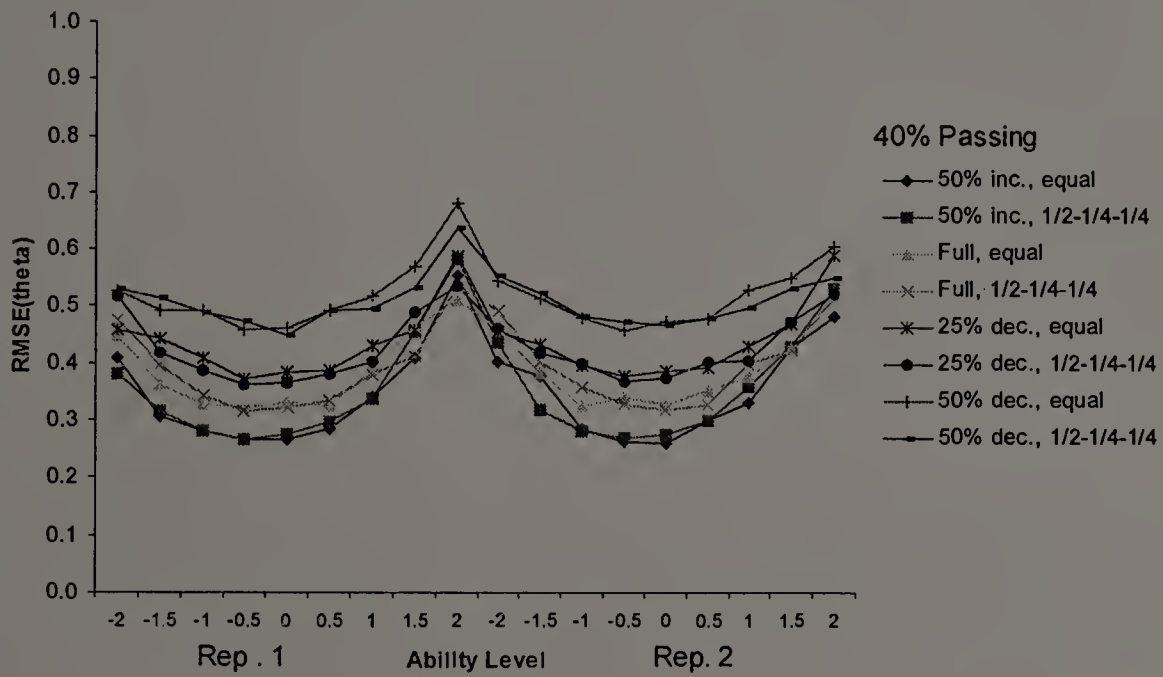
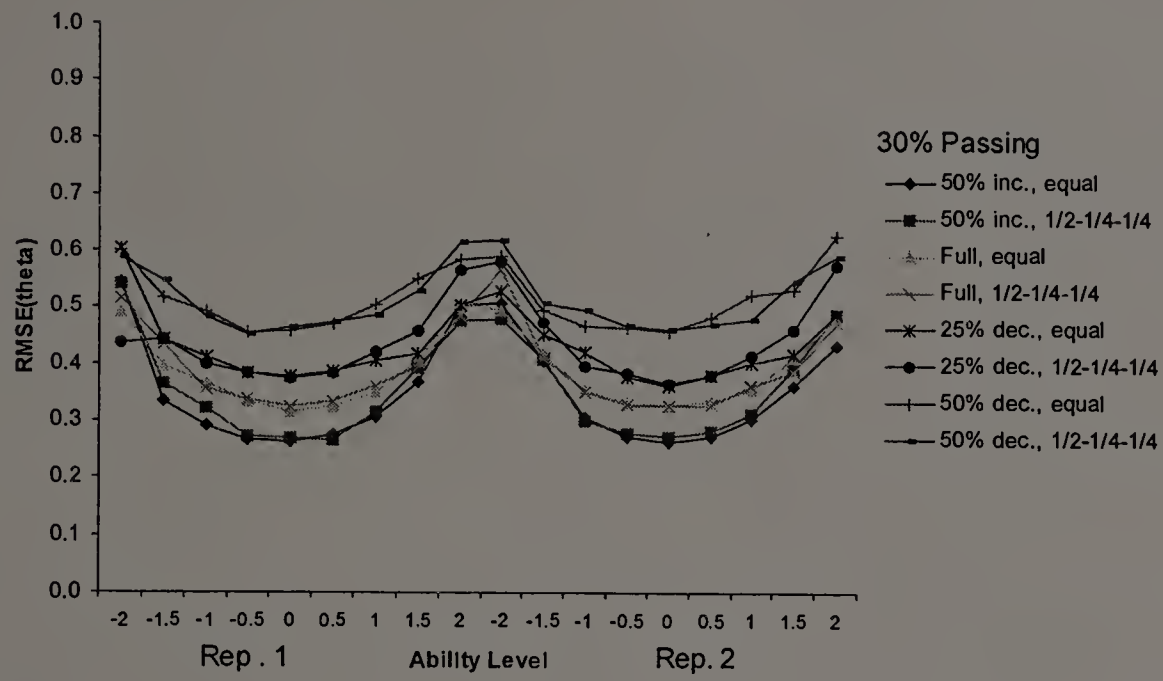


Figure 4.15. RMSEs for Random Routing with 1-2-3 Design at Three Pass Rates

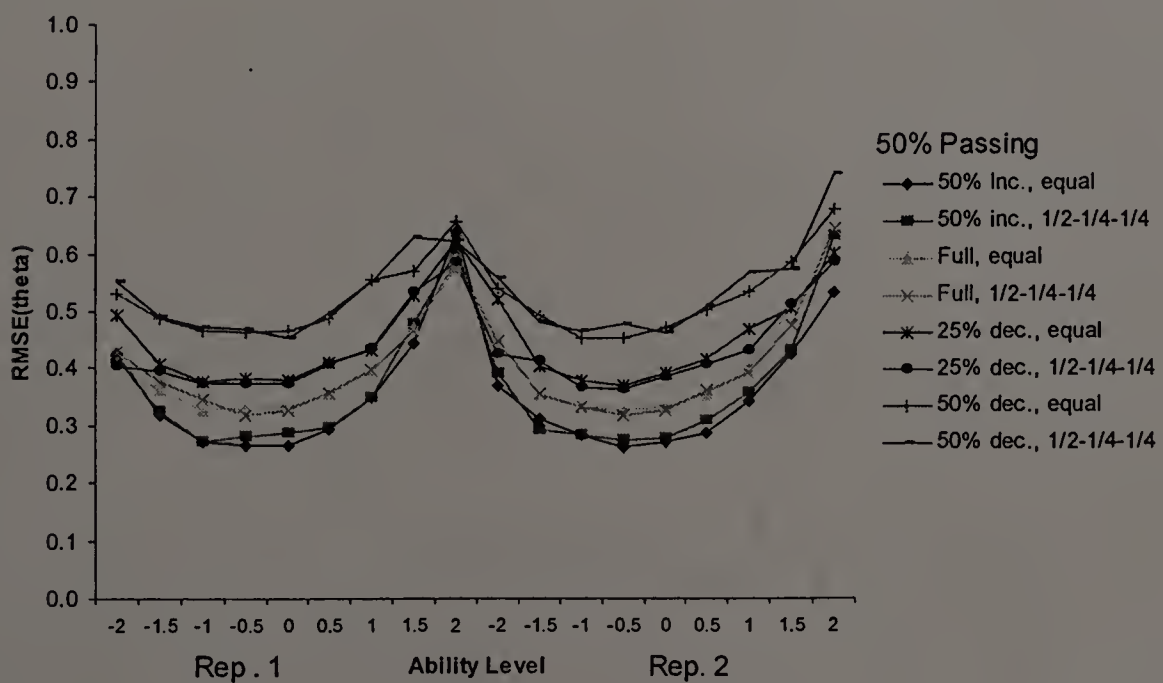
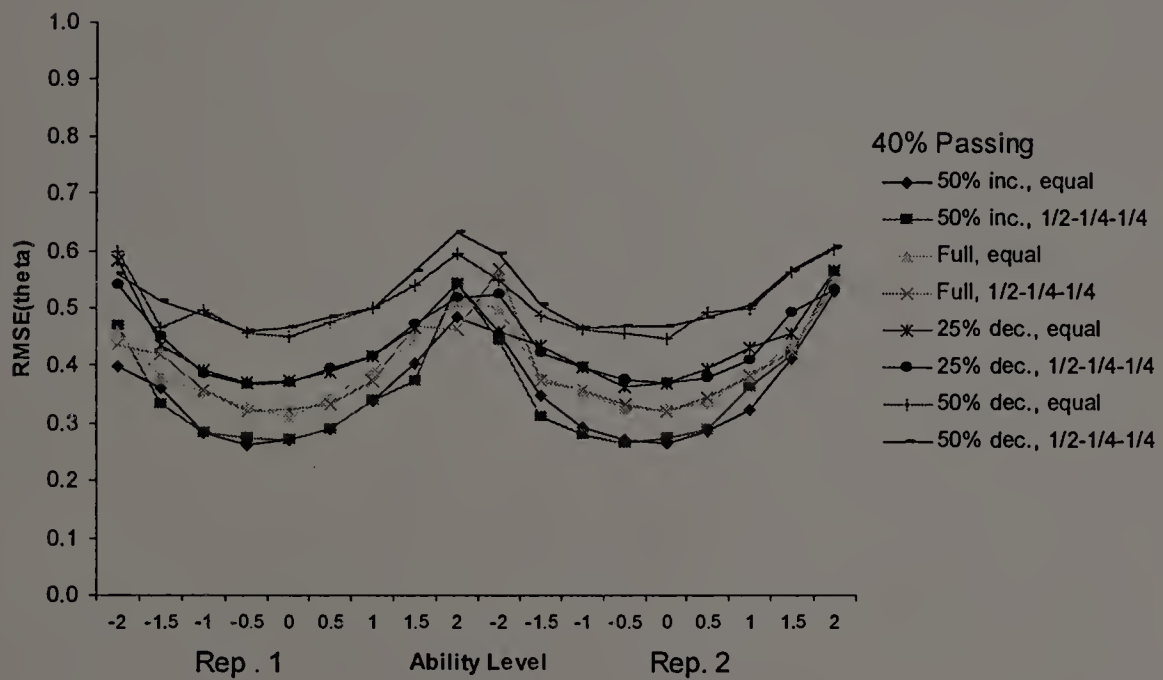
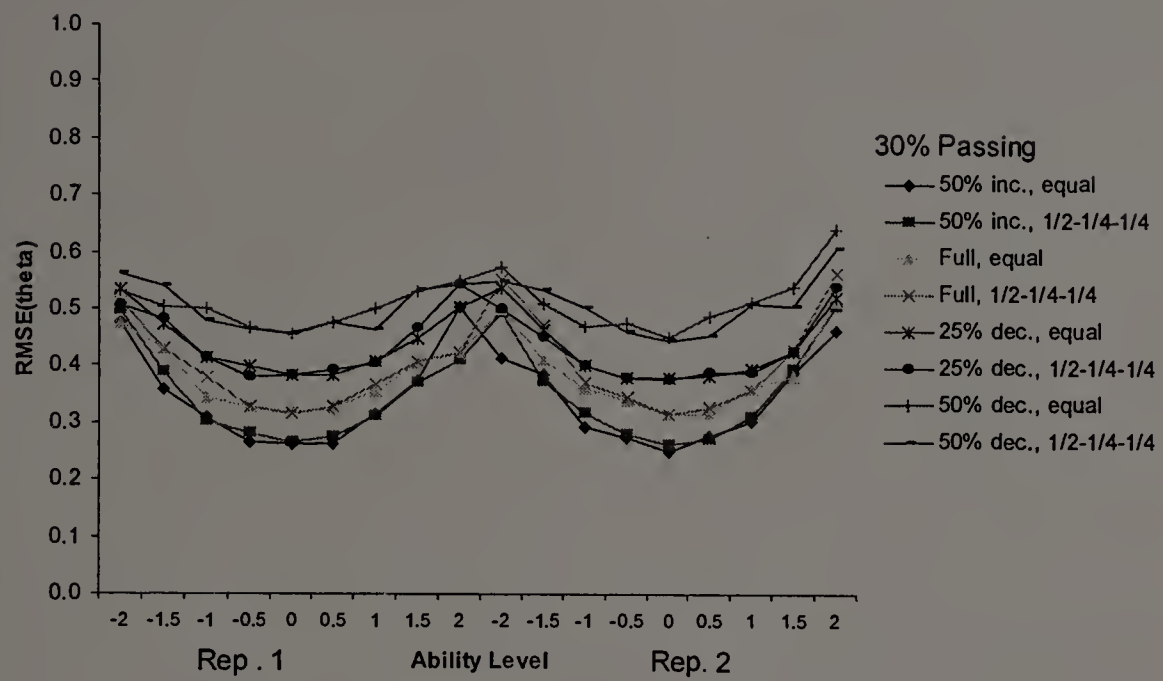
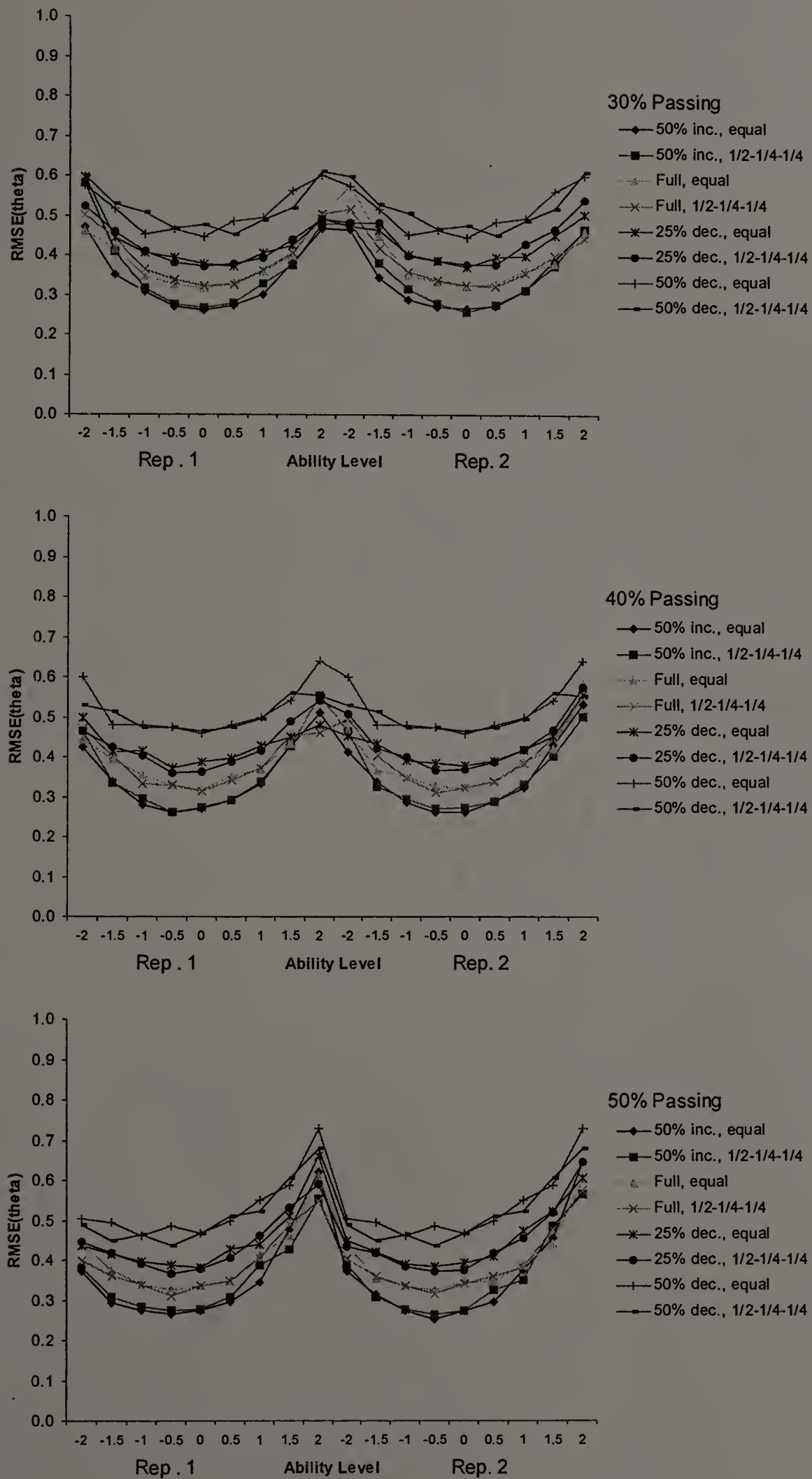


Figure 4.16. RMSEs for Random Routing with 1-3-2 Design at Three Pass Rates



CHAPTER 5

CONCLUSIONS

5.1 Conclusions

Many previous studies have documented the quality of measurement associated with multi-stage tests relative to other test designs. The current simulation study was carried out to help practitioners understand better some of the psychometric properties of multi-stage tests because there are many design variables to consider in constructing and using such tests.

Fixed in this study was the total number of stages, the number of items per module, and the total test length. Of course these are important design considerations as well, but these variables were fixed to be consistent with typical values while exploring the effects of other, less-well-understood variables. It would be practically impossible to study all interest design variables simultaneously. The variables of interest here included the total amount of test information (4 levels), the distribution of test information across stages of the MST (2 levels), the choice of design structure (4 levels), and the routing rules implemented to move candidates from stage to stage (4 levels). Each combination of variables was considered at three levels of passing rates. The result was a study involving 384 conditions ($4 \times 2 \times 4 \times 4 \times 3$).

Results of interest in this study were selected to reflect those of importance in credentialing and licensure assessment: decision accuracy, decision consistency, and ability estimation. In addition, as a practical concern, the proportion of candidates being routed to each possible path in each design structure was also evaluated to inform test development with regard to module exposure rates.

The candidate population and the test characteristics implemented were designed to closely emulate the circumstances of a large-scale, high-stakes credentialing examination. The item bank from which modules were assembled was constructed based on actual item parameters from previous forms of a high-stakes credentialing exam, and the test information function too was based on these forms.

The multi-stage test design, as implemented in this study, clearly provided highly reliable and accurate results with respect to both ability estimation and pass-fail decisions. Overall, the RMSEs, kappa values, and levels of decision consistency and accuracy represent the 'best-case' statistics that would be seen with the use of multi-stage tests, as model-data fit is high. Nevertheless, the results should be illuminating to those with an interest in MST and how such tests might be designed in practice.

The results from this study concerning test information have particular relevance for credentialing agencies. As a general rule, high levels of test information provide better measurement than lesser levels, and certainly this pattern of results was observed in this study. However, it was an interaction between levels of test information and the division of test information across the stages of an MST that emerged as a particularly interesting finding. More information overall in the test lent itself to an equal division of information across stages, while in conditions with less test information comparable measurement results were observed to be associated with a strategy where half of the test information was collected at Stage 1 and one-quarter of the total test information was gathered in the two subsequent stages. Basically, the finding seemed to be that relatively low levels of information should be avoided at Stage 1.

This finding has significance for testing agencies. First, a strategy that allows for comparable measurement results to be obtained with less test information may be quite desirable to testing agencies. A test information function can be met in two ways, either with regard to the 1) quality or 2) number of items. Thus, when it is possible to use a lower test information function, that can be accomplished by using fewer but higher-quality items or more items and drawing more fully from the quality range in the item bank. The second meaning of this result is that it suggests that employing unbalanced levels of test information across stages may well be beneficial for testing in some contexts, and sometimes gathering higher levels of test information earlier in the test to make better routing decisions earlier may be helpful. However slight the benefit, any improvement in the accuracy of decision outcomes due to increased efficiency of the routing at earlier points in the test is a highly desirable goal in test development.

The results relating to the routing rules implemented were likewise interesting, and have implications for implementing multi-stage tests. One strategy, the Random approach, did not take examinee ability into account whatsoever, and measurement and decision results across candidates for this method were lower – but not substantially so – than the other methods that did use ability estimates or number-correct scores in making routing decisions. But it is not likely that this result is generalizable. Were the modules to be positioned further apart within stages, measurement results from the Random method would likely be poorer than evidenced here.

Among the strategies that did incorporate estimates of ability into the routing decisions made, the DPI method did give results that were slightly poorer than the Proximity and Number-Correct methods. Almost certainly, this finding is due to the fact

that the DPI method was primarily focused on equalizing the distribution of candidates across modules and not specifically matching candidates to modules where test information was optimal for the ability distribution.

At the same time results obtained by the DPI methods were only slightly less accurate and consistent than using the Proximity or Number-Correct strategies, and these latter two approaches were highly consistent with one another. From the perspective of a test developer, all these methods are comparable in complexity to implement, and so the choice of strategy does remain one driven by measurement concerns. All things considered, the logic of the Proximity method may be considered to be the most appropriate and defensible of the four methods for high-stakes decisions, as it involves assigning candidates to the module empirically determined to most nearly match their estimated ability.

No differences in measurement outcomes of interest were detected with respect to the choice of design strategy employed. As testing agencies consider the merits of different module configurations, in terms of outcomes no differences due to using two or three modules at stages 2 and 3 were found. However, the decision to limit the variation in difficulty of second and third stage modules may have limited the extent to which the test designs produced different results.

Concerning design strategy, in the absence of clear measurement advantages, the bigger operational concern for programs seems to be using more than two stages so that the candidates do not have the perception of being unable to pass if they do poorly at Stage 1. Operationally, there may be certain benefits to only having to manage one cut-score in moving from stage to stage, and concentrating resources on making sure that the

routing based on that single cut between each stage is as precise as possible. For this reason, some test developers may prefer the 1-2-2 design given that its use does not result in any lowering of DA and DC.

In comparing results across several different pass rates (30%, 40%, 50%), the decrease in accuracy and consistency observed as the passing score was moved from +.521 to .000 was clear. At the same time, this decrease was not so striking that practitioners would expect wildly different results depending on the placement of the passing score. The results from this study are likely to be generalizable to passing scores set that results in pass rates ranging from at least 30% to 70%.

The results from this study suggest that the design overall does provide a high level of measurement quality for a variety of implementation structures and strategies. But in simulation studies like this one using model-generated data, model-data fit is high and findings do tend to over-predict the findings observed in practice.

Ultimately, the findings highlighted here do represent an investigation to clarify certain specific aspects of an under-researched approach to adapting tests to examinees in the specific context of credentialing testing where decisions are the overriding concern. However, the inclusion of measurement accuracy as an outcome of interest further generalizes the conclusions to other testing contexts and test uses, where accuracy of ability estimation itself is desired. These conclusions suggest that some MST design variables do not significantly shape the results (module arrangement being a prime example of this) but the relationship between measurement outcomes and other design variables (such as amount of test information and the stages where that information is

collected) is more complex and test developers should weigh such decisions carefully in light of findings from this study.

Generalizing findings from simulation studies is always problematic. In practice, psychometric models never completely explain candidate performance, and with the MST design, there is always the potential psychological impact on candidates if they notice a shift in test difficulty. At the same time, two findings seem to stand out in this research: (1) with limited amounts of overall test information, it may be best to capitalize on the information that is available with accurate branching decisions at Stage 1, and (2) unless for reasons of content validity, or to convince candidates they have been rigorously assessed, there may be little advantage of exceeding test information much above 10 because the gains in decision consistency and decision accuracy appear to be quite small.

5.2 Directions for Future Research

There are a number of research questions that seem worthy of follow-up research. The first direction of interest concerns further investigation of various routing strategies in the context of considerably shortened tests. As mentioned above, a probable cause for the results by routing strategy in this study not being more distinct is that after 60 items' worth of testing, any differences due to poorer or better routing may well be rendered less evident than they would be with shorter tests. The interesting finding in this study, with tests of lower test information function compared to the baseline tests, provides an indication of the impact of shortening tests. It would be interesting to vary the lengths of the individual stage-level tests in the context of generally shorter tests, say 30 or 40

items. When test (and accordingly, module) lengths are reduced, more significant differences in the methods may well become clear. In terms of maximally estimating examinee abilities in such high-stakes settings, the value of knowing the preferred method for doing so should not be underestimated.

A similar future investigation might well focus more closely on the splitting of information among stages. Interesting patterns of results were observed in this study. More effective targetting of module information in relation to passing scores and ability distribution would be another question worthy of study.

Another important extension of this study would be to build more error into the simulations and repeat them, to better reflect the kinds of errors that would be seen in practice. Simulation approaches such as adding a second dimension correlated to the first should be considered, since with less good model fit due to the second dimension, it would be possible to obtain simulation results that might better reflect those that could be obtained in practice.

Finally, the branching of many testing programs into measuring skills and abilities that are more complex in nature raises the possibility of multi-stage tests using polytomously-scored tasks. Here, a stage might consist of two or more polytomously-scored items. Utilizing the adaptive structure of multi-stage tests to improve measurement precision by improving selection of such items for administration as part of a stage-based test structure may well be a direction of interest for researchers. Indeed, approaches using polytomous items in MST could explore the efficacy of both dichotomous and polytomous items or polytomous items alone.

BIBLIOGRAPHY

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement*, 27(3), 241-253.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Angoff, W., & Huddleston, E. (1958). *The multi-level experiment: a study of a two-level testing system for the College Board Scholastic Aptitude Test*. (Statistical report No. SR-58-21). Princeton, New Jersey: Educational Testing Service.

Armstrong, R., Jones, D., Koppel, N., & Pashley, P. (2000, April). *Computerized adaptive testing with multiple forms structures*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bayroff, A. G., & Seeley, L. C. (1967, June). *An exploratory study of branching tests* (Technical Report Note 188). Washington, D.C.: U. S. Army Behavioral Science Research Laboratory.

Bayroff, A. G., Ross, R. M., & Fischel, M. A. (1974, December). *Development of a programmed testing system*. (Technical Paper 259). Arlington, VA: U. S. Army Research Institute.

Berger, M. P. F. (1994). A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement*, 18(2), 141-153.

Bergstrom, B. A., & Lunz, M. E. (1992). Confidence of pass/fail decisions for computer-adaptive and paper-and-pencil examinations. *Evaluation and the Health Professions*, 15(4), 453-464.

Betz, N., & Weiss, D. (1973). *An empirical study of computer adaptive two-stage ability testing* (Research Report No. 73-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N., & Weiss, D. (1974). *Simulation studies of two-stage testing* (Research Report No. 74-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Method program.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for the diagnosis of the intellectual level of abnormals]. *LiAnnée Psychologique*, 11, 191-245.

Binet, A., & Simon, T. (1908). Le développement de l'intelligence chez les enfants [The development of intelligence in children]. *L'Année Psychologique*, 14, 1-90.

Bock, R. D., & Mislevy, R. (1988). Comprehensive educational assessment for the states: The duplex design. *Educational Evaluation and Policy Analysis*, 10, 89-105.

Bock, R. D., & M. F. Zimowski. (1989). *The duplex design: Giving students a stake in educational assessment*. Chicago: National Opinion Research Center, Methodology Research Center.

Bock, R. D., & Zimowski, M. F. (1998). Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP. *NAEP Validity Studies*. Palo Alto, CA: American Institutes for Research.

Boughton, K. A., & Gierl, M. J. (2000, April). *Automated test assembly procedures for criterion-referenced testing using optimisation heuristics*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153-168.

Carey, P. A. (1999, April). *The use of linear-on-the-fly testing for TOEFL reading*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Castle, R. A. (1997). *The relative efficiency of two-stage testing versus traditional multiple choice testing using item response theory in licensure*. Unpublished doctoral dissertation, University of Nebraska.

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

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

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd Ed.). Urbana: University of Illinois Press.

Dodd, B. G., & Fitzpatrick, S. J. (2002). Alternatives for scoring CBTs. In C. N. Mills, M. T. Potenza, J.J. Fremer, and W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* [pp. 215-236]. Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

Du, Y., Lewis, C., & Pashley, P. J. (1993). Computerized mastery testing using fuzzy set decision theory. *Applied Measurement in Education*, 6, 181-193.

Ferguson, R. L. (1969a). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh, PA: University of Pittsburgh Learning and Research Development Center. (ERIC Document Reproduction Service No. ED 037 089).

Ferguson, R. L. (1969b). *The development, implementation, and evaluation of a computer assisted branched test for individually prescribed instruction*. Unpublished doctoral dissertation. University of Pittsburgh, Pittsburgh, PA. (University Microfilms No. 70-4530).

Folk, V. G., & Smith, R. L. (2002). Models for delivery of computer-based tests. In C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* [pp. 41-66]. Mahwah, NJ: Lawrence Erlbaum Associates.

Foong, Y.-Y., & Lam, T.-L. (1991, April). *Development and evaluation of hierarchical testlets in two-stage tests using integer linear programming*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

Green, B. F., Jr. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. [pp. 69-80]. Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Jr., Bock, R. D., Humphreys, L. G., Linn, R. B., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.

Hambleton, R. K. (2002a). New CBT technical issues: developing items, pretesting test security, and item exposure. In C. N. Mills, M. T. Potenza, J.J. Fremer, and W. C. Ward (Eds.), *Computer-based testing; Building the foundation for future assessments* [pp.193-204]. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. (2002b, April). *Test design*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., & Xing, D. (2002). *Comparative analysis of optimal and non-optimal computer-based test designs for making pass-fail decisions*. Paper presented at the annual meeting of the Canadian Educational Research Association, Toronto.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* [pp. 341-366]. Boston: Kluwer Academic Publishers.

Jodoin, M. G. (2003). *MSTSIM5* [Computer software]. Amherst, MA: University of Massachusetts, School of Education.

Jodoin, M. G. (2002, June). *Reliability and decision accuracy of linear parallel form and multi stage tests with realistic and ideal item pools*. Paper presented at the International Conference on Computer-Based Testing and the Internet, Winchester, England.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kim, H. (1993). *Monte Carlo simulation comparison of two-stage testing and computer adaptive testing*. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kim, H., & Plake, B. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education. Atlanta, GA.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* [pp.257-283]. New York: Academic Press.

Krathwohl, D. R., & Huyset, R. J. (1956). The sequential item test (SIT). *American Psychologist*, 2, 419.

Lam, T.-L., & Foong, Y.-Y. (1991, April). *Development and evaluation of hierarchical testlets in two-stage tests using integer linear programming*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Larkin, R., & Weiss, D. (1975). *An empirical comparison of two-stage and pyramidal adaptive testing* (Research Report No. 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Lee, G., & Frisbie, D.A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12, 237-255.

Lewis, C., & Sheehan, K. (1988). Computerized mastery testing. *Machine-Mediated Learning*, 2, 283-286.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 367-386.

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

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* [pp. 139-183]. New York: Harper and Row.

Lord, F. M. (1971a). Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1), 3-31.

Lord, F. M. (1971b). The self-scoring flexi-level test. *Journal of Educational Measurement*, 8, 147-151.

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

Lord, F. M. (1971d). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66, 707-711.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test*. Research Bulletin 74-30. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14(2), 227-238.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loyd, B. H. (1984). *Efficiency and precision in two-stage adaptive testing*. Paper presented at the Annual Meeting of the Eastern Educational Research Association, West Palm Beach, FL.

Luecht, R. M. (1997, March). *An adaptive sequential paradigm for managing multidimensional content*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (1998). *CASTISEL* [Computer software]. Philadelphia, PA: National Board of Medical Examiners.

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

Luecht, R., Brumfield, T., & Breithaupt, K. (2002, April). *A testlet-assembly design for the Uniform CPA Exam*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computer adaptive test administration formats. *Journal of Educational Measurement*, 31(3), 251-263.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14(2), 117-138.

Mills, C. N., Potenza, M. T., Fremer, J.J., and Ward, W. C. (Eds.). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mills, C.N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9, 287-304.

Papanastasiou, E. C. (2002, April). *A rearrangement procedure for administering adaptive tests with review options*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Patelis, T. (2000, April). *An overview of computer-based testing*. New York, NY: The College Board Office of Research and Development Research Notes (RN-09).

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

Patsula, L. N., & Hambleton, R. K. (1999, April). *A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing*. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec.

Patterson, J. J. (1962). *An evaluation of the sequential method of psychological testing*. Unpublished doctoral dissertation, Michigan State University.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* [pp. 237-255]. New York: Academic Press.

Reese, L.M., & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing*. (Computerized Testing Report 96-04). Newtown, PA: Law School Admissions Council.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design*. (Law School Admissions Council Computerized Testing Report 97-02). Newtown: PA: Law School Admissions Council.

Rock, D. A., Pollack, J. M., and Quinn, P. (1995, August). *Psychometric report for the NELS:88 base year through second follow-up*. Washington, D.C.: Office of Educational Research and Improvement, U.S. Department of Education, NCES 95-382.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2003, April). A comparison of multi-stage tests with computerized adaptive and paper & pencil tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17*.

Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing*. (Law School Admissions Council Computerized Testing Report 97-01). Newtown: PA: Law School Admissions Council.

Sheehan, K., & Lewis, C. (1992). Computerized master testing with non-equivalent testlets. *Applied Psychological Measurement, 16*, 65-76.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247.

Smith, R. L., & Lewis, C. (1995, April). *A Bayesian computerized mastery model with multiple cut scores*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Smith, R. L., & Lewis, C. (1998, April). *Expected losses for individuals in computer mastery testing*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Smith, R. L., & Lewis, C. (2002, April). *A comparison of computer mastery testing models when pool characteristics vary*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education*, 7, 211-222.

Thissen, D. (1998, April). *Scaled scores for CATs based on linear combinations of testlet scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen, *Computerized adaptive testing: A primer* (2nd Ed.) [pp. 101-132]. Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247-260.

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement*, 24(3), 225-240.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, 35(3), 185-198.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computer-adaptive, and self-adaptive testing. *Applied Measurement in Education*, 7, 53-79.

Vos, H. J. (2000a). A Bayesian procedure in the context of sequential mastery testing. *Psicológica*, 21, 191-211.

Vos, H. J. (2000b, April). *Adaptive mastery testing using a multidimensional IRT model and Bayesian sequential decision theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Vos, H. J., & Glas, C. A. W. (2000). Testlet-based adaptive mastery testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* [pp. 289-310]. Boston, MA: Kluwer Academic Publishers.

Vos, H. J., & Glas, C. A. W. (2001, April). *Multidimensional IRT based adaptive sequential mastery testing*. Paper presented at the annual meeting of the National Council in Measurement in Education, Seattle, WA.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12(1), 15-20

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, 8(2), 157-186.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* [pp. 245-270]. Boston, MA: Kluwer Academic Publishers.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 27, 1-14.

Wainer, H., Sireci, S., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, 28, 197-219.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wightman, L. F. (1998). Practical issues in computerized test assembly. *Applied Psychological Measurement*, 22, 292-302.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Wise, S. L. (1999b, April). *Comparison of stratum scored and maximum-likelihood scored CATs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135-155.

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

Xing, D., & Hambleton, R. K. (2001, April). *Impact of several computer-based testing variables on the psychometric properties of credentialing exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Xing, D., & Hambleton, R. K. (2002, April). *Impact of item quality and item bank size on the psychometric properties of computer-based credentialing exams*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Yi, Q., Hanson, B. A., Widiatmo, H., & Harris, D. J. (2001, April). *Comparison of the SPRT and CMT procedures in computerized classification tests*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Zenisky, A. L. (2002). *An empirical investigation of selected multi-stage testing design variables on test assembly and decision accuracy outcomes for credentialing exams*. Center for Educational Assessment Research Report No. 469. Amherst, MA: University of Massachusetts, School of Education

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 1-16.

Zimowski, M. F. (1988, April). *The duplex design: An evaluation of the two-stage testing procedure*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zimowski, M. F. (1989.) *Advantages of two-stage testing in student-reporting assessments*. Paper presented at the annual meeting of the Education Commission of the States, Boulder, CO.

