The Dissertation Committee for Meredith Sibley Butterfield certifies that this is the approved version of the following dissertation:

Comparing Item Selection Methods in Computerized Adaptive Testing Using the Rating Scale Model

Committee:

____

Barbara G. Dodd, Supervisor

____

Tiffany A. Whittaker

____

Jodi M. Casabianca-Marshall

____

Matthew Hersh

Comparing Item Selection Methods
in Computerized Adaptive Testing
Using the Rating Scale Model

by

Meredith Sibley Butterfield, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2016

Dedication

For my children
Owen and Townsend Butterfield

Acknowledgements

I am deeply grateful to my dissertation advisor, Barbara Dodd. Through her courses and as her teaching assistant, I acquired a thorough understanding of psychometrics. Dr. Dodd has been generous with her time, guidance, and encouragement over the years of my doctoral studies, during my qualifying process, and ultimately through this dissertation study.

I would also like to thank my other committee members, Dr. Tiffany Whittaker, Dr. Jodi Casabianca-Marshall and Dr. Matt Hersh for their suggestions as well as their courses over the years of my studies. Each person has provided valuable support and feedback. Their teaching and insightful suggestions have enhanced my knowledge of the field of quantitative methods and the quality of my study.

To my parents, Bob and Sue Sibley, I was fortunate to always have your love, support, and encouragement, especially in academic endeavors. Throughout my childhood, you pushed me to meet my potential and modeled hard work and intellectual curiosity.

I could not have completed my studies and this dissertation without the support and love of my husband, Brian. Thank you for supporting my decision and making it possible for me to return to school even with a new baby at home, and then your unwavering encouragement and help through baby number two. I love you.

Comparing Item Selection Methods in Computerized Adaptive Testing Using the Rating Scale Model

Meredith Sibley Butterfield, PhD.
The University of Texas at Austin, 2016

Supervisor: Barbara G. Dodd

Computer Adaptive Testing (CAT), a form of computer-based testing that selects and administers items that match the examinee's trait levels, can be shorter in length and maintain comparable or greater measurement precision than traditional fixed-length paper-and-pencil testing. Administration of computer-based patient reported outcome (PRO) measures has increased recently in the medical field. Because PRO measures often have small item pools, small numbers of items administered, and populations in poor health, the benefits of CATs are especially advantageous. In CAT, Maximum Fisher information (MFI) is the most commonly used item selection procedure since it is easy to use and computationally simple. However, its main drawback is the attenuation paradox. If the estimated trait level of the examinee is not the true trait level, the items selected will not maximize information at the true trait level and the measurement is less precise. To address this issue, alternative item selections methods have been proposed. In studies, these alternatives have not performed better than MFI. Recently, Gradual Maximum Information Ratio (GMIR) item selection method was proposed and previous findings suggest GMIR could be beneficial for a short CAT.

This simulation study compared GMIR and MFI item selection methods under conditions specific to the constraints of the PRO measures. GMIR and MFI are compared under Andrich's Rating Scale Model (ARSM) across two polytomous item pool sizes (41 and 82), two population latent trait distributions (normal and negatively skewed), and three combination maximum number of item and minimum standard error stopping rules (5/0.54, 7/0.46, 9/0.40). The conditions were fully crossed. Performance was evaluated in terms of descriptive statistics of the final trait estimates, measurement precision, conditional measurement precision, and administration efficiency. Results found GMIR had better measurement precision when the test length was 5 items, with higher mean correlations between known and estimated trait levels, smaller mean bias, and smaller mean RMSE. An effect of item pool size and population latent trait distribution was not found. Across item selection methods, measurement precision increased as the test length increase, but with diminishing returns from 7 to 9 items.

Table of Contents

List of Tables ix

List of Figures x

Chapter 1 Introduction 1

Chapter 2 Literature Review 8

Item Response Theory 8

Dichotomous IRT Models 9

Polytomous IRT Models 13

Assumptions 18

Item and Test Information 19

Computer Adaptive Testing 20

Advantage of CAT 21

Item Pool 23

Item Selection 27

Trait Estimation 42

Stopping Rule 45

Statement of Problem 48

Chapter 3 Methodology 51

Design Overview 51

Item Pool and Test Characteristics 52

Data Generation 54

CAT Simulation 56

Data Analysis 58

Chapter 4 Results 61

Nonconvergent Cases 61

Estimated Thetas 65

Overall Measurement Precision 68

Conditional Measurement Precision 73

Test Efficiency 135

Chapter 5 Discussion 137

Research Questions 137

Practical Applications 143
Limitations and Future Research 145
Appendix A Plots of Bias, RMSE, and SE Conditional on Item Number 147
References 166

List of Tables

Table 1. Descriptive Statistics for Item Parameter Estimates for Shoulder Functioning Data........53

Table 2. Nonconvergent Cases Averaged across 100 Replications........64

Table 3. Estimated Thetas and Standard Error Descriptive Statistics Averaged across 100 Replications........67

Table 4. Correlation Coefficients between Known and Estimated Thetas Averaged across 100 Replications........71

Table 5. Bias and RMSE Averaged across 100 Replications........72

Table 6. Number of Items Administered Descriptive Statistics Averaged across 100 Replications........136

List of Figures

Figure 1. ICCs for 1PL, 2PL, and 3PL Models..12

Figure 2. CRCs for Five-category Item 17

Figure 3. Information Function for Shoulder Functioning Scale 82-Item Item Pool Calibrated According to Andrich's Rating Scale Model...................54

Figure 4. Negatively Skewed Theta Distribution for One Sample............................55

Figure 5A. Plots of Mean Bias Conditional on Known Theta, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 75

Figure 5B. Plots of Mean Bias Conditional on Known Theta, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 76

Figure 5C. Plots of Mean Bias Conditional on Known Theta, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 77

Figure 6A. Plots of Mean Bias Conditional on Known Theta, the 5 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 78

Figure 6B. Plots of Mean Bias Conditional on Known Theta, the 7 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 79

Figure 6C. Plots of Mean Bias Conditional on Known Theta, the 9 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 80

Figure 7A. Plots of Mean RMSE Conditional on Known Theta, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 81

Figure 7B. Plots of Mean RMSE Conditional on Known Theta, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 82

Figure 7C. Plots of Mean RMSE Conditional on Known Theta, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 83

Figure 8A. Plots of Mean RMSE Conditional on Known Theta, the 5 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 84

Figure 8B. Plots of Mean RMSE Conditional on Known Theta, the 7 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 85

Figure 8C. Plots of Mean RMSE Conditional on Known Theta, the 9 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 86

Figure 9A. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 87

Figure 9B. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 88

Figure 9C. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 89

Figure 10A. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 5 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 90

Figure 10B. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 7 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 91

Figure 10C. Plots of Mean Standard Error (SE) Conditional on Known Theta, the 9 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 92

Figure 11A. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Population Conditions....95

Figure 11B. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Population Conditions....96

Figure 11C. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Population Conditions....97

Figure 12A. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 5 Item Stopping Rule, and Normally Distributed Population Conditions....98

Figure 12B. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 7 Item Stopping Rule, and Normally Distributed Population Conditions....99

Figure 12C. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 100

Figure 13A. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 101

Figure 13B. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 102

Figure 13C. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 103

Figure 14A. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 104

Figure 14B. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 105

Figure 14C. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 106

Figure 15. Plot of Mean Bias Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool 107

Figure 16A. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 110

Figure 16B. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 111

Figure 16C. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 112

Figure 17A. Plots of Mean RMSE Conditional on Item Number for Known Theta =

-1, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 113

Figure 17B. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 114

Figure 17C. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 115

Figure 18A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 116

Figure 18B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 117

Figure 18C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 118

Figure 19A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 119

Figure 19B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 120

Figure 19C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 121

Figure 20A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 122

Figure 20B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 123

Figure 20C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 124

Figure 21A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 5 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 125

Figure 21B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 7 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 126

Figure 21C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, and Negatively Skewed Distributed Population Conditions 127

Figure 22A. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Population Conditions 130

Figure 22B. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Population Conditions 131

Figure 22C. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Population Conditions 132

Figure 23. Plot of Mean SE Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, Normally Distributed Population, and 41 Item Pool 133

Figure 24. Plot of Mean SE Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, Normally Distributed Population, and 41 Item Pool 133

Figure 25. Plot of Mean SE Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool 134

Figure 26. Plot of Mean SE Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool 135

Figure A.1. Plots of Mean Bias Conditional on Item Number for Known Theta = -2 for Negatively Skewed Population Distribution........ 147

Figure A.2. Plots of Mean Bias Conditional on Item Number for Known Theta = -1 for Negatively Skewed Population Distribution........ 148

Figure A.3. Plots of Mean Bias Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distribution..149

Figure A.4. Plots of Mean Bias Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distribution..150

Figure A.5. Plots of Mean Bias Conditional on Item Number for Known Theta = 2 for Normally Distributed Population 151

Figure A.6. Plots of Mean Bias Conditional on Item Number for Known Theta = 2 for Negatively Skewed Population Distribution..152

Figure A.7. Plots of Mean RMSE Conditional on Item Number for Known Theta =-2 for Negatively Skewed Population Distribution..153

Figure A.8. Plots of Mean RMSE Conditional on Item Number for Known Theta =-1 for Negatively Skewed Population Distribution..154

Figure A.9. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distribution..155

Figure A10. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distribution..156

Figure A.11. Plots of Mean SE Conditional on Item Number for Known Theta = -2 for Negatively Skewed Population Distribution..157

Figure A.12. Plots of Mean SE Conditional on Item Number for Known Theta = -1 for Normally Distributed Population 158

Figure A.13. Plots of Mean SE Conditional on Item Number for Known Theta = -1 for Negatively Skewed Population Distribution..159

Figure A.14. Plots of Mean SE Conditional on Item Number for Known Theta = 0 for Normally Distributed Population 160

Figure A.15. Plots of Mean SE Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distribution..161

Figure A.16. Plots of Mean SE Conditional on Item Number for Known Theta = 1 for Normally Distributed Population 162

Figure A.17. Plots of Mean SE Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distribution..163

Figure A.18. Plots of Mean SE Conditional on Item Number for Known Theta = 2 for Normally Distributed Population 164

Figure A.19. Plots of Mean SE Conditional on Item Number for Known Theta = 2 for Negatively Skewed Population Distribution 165

Chapter 1: Introduction

Computer Adaptive Testing (CAT) is a form of computer-based testing that adapts to the examinee. Instead of administering an identical set of items to every examinee like traditional fixed-length paper-and-pencil testing (P&P), the CAT program selects and administers items that match the examinee's trait levels, so every examinee receives a unique set of items. If the examinee answers an item incorrectly, which indicates the item was too difficult, the computer will administer an easier item next. Conversely, if an examinee answers an item correctly, the computer then administers a more difficult item.

Item response theory (IRT) is the psychometric approach on which computer adaptive testing is based. IRT uses item properties or parameters and examinee's item responses to estimate examinee's traits. IRT allows a person's trait measurement to only vary within a linear transformation, regardless of the level of the trait and regardless of the item administered. A CAT can administer only a subset of items from a larger item pool. IRT models predict the probability of a response to an item, conditional on the person and item parameters. When the response options are binary, dichotomous models are appropriate, and when there are more than two response options, polytomous models are appropriate. Common dichotomous models include the one, two, and three parameter logistic models (Birnbaum, 1968; Rasch, 1960). Examples of polytomous models include the Graded Response Model (Samejima, 1969), the Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), and Andrich's Rating Scale Model (1978).

CAT is used in a number of fields including education, business, military, and medicine. Examples of CATs are the Graduate Record Examination (GRE), a number of Microsoft Certified Professional exams, the Armed Services Vocational Aptitude Test Battery (ASVAB), and the Patient-Reported Outcomes Measurement Information System (PROMIS) Health Assessment Questionnaire Disability Index (HAQ). Recently, in the health and medical fields, administration of computer-based patient reported outcome (PRO) measures has increased and can be expected to grow. Ware, Bjorner and Kosinski (2000) stated "the health case industry needs more practical tools to monitor population health on a large scale as well as more precise tools to identify those who need and are most likely to benefit from treatment." The National Institutes of Health (NIH) Roadmap Initiative is making a systematic effort to develop and make available on a public web-based system thousands of items and CAT systems for a large number of PRO assessment measures, known as PROMIS (Reeve, 2006). The NIH is looking to solve the problems of small clinical trials with small individual labs and small sample sizes. Jette and Haley (2005) complain of the "data incompatibility across instruments" and the "inability of different outcome assessments to talk to each other" as a limitation of current methodology. Reeve (2006) states that the goals of PROMIS are to electronically administer individually tailored PRO measures, collect the PRO data for research, and provide immediate health reports to the patient, health care provider, and/ or researcher. Since secrecy of items is not necessary in the health care industry as it is in educational testing, collaboration and open sharing of items is easier. Walker , Böhnke, Cerny and Strasser (2010) stated that PROMIS "aims to revolutionize the way (PRO) tools are selected and used in clinical research and practice evaluation." The PROMIS Assessment Center contains item pools and scales in a number of domains: global health, mental health (e.g. alcohol use), physical health (e.g. pain interference), and social health (e.g. satisfaction with social roles and activities)(*PROMIS Full Domain Framework*, 2016). CATs have been developed that evaluate health outcomes in a variety of areas. CATs have been developed to assess headache impact (Ware et al., 2003), depression (Fliege et al., 2005), anxiety (Walter et al., 2007), and many other health issues. Researchers in the medical field have studied the use of CAT for patient-based health status measures and have found the use of computer adaptive testing to be advantageous. Reeves stated in 2006 that it is "the brink of a new era for health outcomes measurement with the availability of CAT-based tools which integrate the advances of modern computer technology and the strengths of modern measurement theory."

When compared to traditional testing, CAT has a number of benefits. Many of these benefits are specifically advantageous to the health care PRO measures due to unique characteristics of the PRO measures as compared to educational testing. While educational testing measures can have item pools in the 100s, PRO measure item banks are generally 50 items (Walker et al., 2010). CAT patient reported outcome measures in current research, in contrast to lengthier educational measures which are usually 20 items or more, often limit the number of items to five to ten (Cook et al., 2008; Cook et al., 2007; Fliege et al., 2005; Ware et al., 2005). The health care field has a different population than the educational testing field; patients are often ill or have physical limitations. It is important for the medical field to understand how a CAT performs under these specific conditions. The primary focus of this study is the application to the specific constraints of and orientation to the PRO measures.

Advantageously, the adaptive testing process increases the efficiency of the test; CATs can be shorter in length and maintain comparable or greater measurement precision. (Cook et al., 2008; Dodd, de Ayala, & Koch, 1995; Embretson & Reise, 2000; Fliege et al., 2005; Wainer, 2000; Ware et al., 2003; Ware, Gandek, Sinclair, & Bjorner, 2005). Tradition non-adaptive P&P testing administers a larger number of items, many are too high for examinees with low trait values many are too low for examinees with high trait values. In education for example, a test would have questions that are too difficult for a low ability student to solve correctly and questions that a high ability student could easily answer all correctly. This results in an examinee being administered unnecessary items. An unnecessarily long measure that takes needless time and energy can be a nuisance in educational and professional settings, but particularly burdensome for patients in a medical setting. Limiting this testing burden is especially advantageous for PROs when the patient has physical limitations or a serious illness. Physicians may monitor patients through repeated administrations of a PRO at appointments or treatments, for cancer for example, where avoiding lengthy surveys would be benevolent and precise measurement of changes across time are important. Administration of a CAT instead of a P&P patient reported outcome measure allows for so few items to be administered and can limit the testing burden while maintaining or increasing measurement precision. Furthermore, specific conditions of various CAT components may be ideal when such a short test length is administered. Another advantage of CAT that is particularly relevant to the medical field is that the number of patients needed for a clinical trial can be reduced, while maintaining statistical power (Fries, 2006; Fries et al., 2014). This allows health care research trials to proceed with fewer resources and less funding. Computer based testing also benefits the health care industry since the evaluation results are easier than P&P to integrate into patients' electronic records (Walker et al., 2010). Systems like PROMIS can improve patient-doctor communication and health-care provider decision making. Electronic records for measures that can be compared across research and practice settings could allow for better evaluation of the effectiveness of treatments, provider, and organizations, which is of growing interest to administrators. Also, NIH will gain a greater ability to monitor and understand causes in disease progression (Reeve, 2006).

One critical component of CATs that increases the efficiency as compared to P&P testing is the method of item selection. Through the item selection method, the CAT can select a unique set of items from the larger item pool. Maximum Fisher information (MFI) is the most commonly used item selection method for CATs (Lord, 1980). MFI method selects the item that maximizes information, item measurement precision, at a specific trait estimate. While the MFI method is easy to use and computationally simple, its main drawback is the attenuation paradox (Lord & Novick, 1968). If the specific estimated trait level is not the true trait level of the examinee, the items selected do not maximize information at the true trait level. If the items are less optimal, the estimates are inefficient and the measurement is less precise. This is particularly an issue in the first few items of the CAT.

To address this weakness, a number of alternatives to MFI have been developed. Veerkamp and Berger (1997) proposed interval information criterion (IIC) and likelihood weighted information criterion (LWIC), which are extensions of MFI. Chang and Ying (1996) proposed Kullback-Leibler (KL) information as an alternative to Fisher information. Owen (1975) and van der Linden (1998) proposed a number of different Bayesian approaches, which select items based on the prior and posterior distributions of the trait estimates. In studies, these alternatives methods have generally estimated trait levels comparably or more poorly than MFI. Van Rijn, Eggen, Hemker, and Sanders (2002) found little difference between IIC and MFI using the PCM. Veldkamp (2003) found KL and IIC performance compared to MFI in terms of mean squared error and overlap of items administered, using the GPCM. Lima Passos et al. (2007) found KL performed comparably to MFI and IIC fluctuated more than MFI in terms of RMSE and bias. Choi and Swartz (2009) found MFI, LWI, and Bayesian methods performed similarly in terms of bias, Root Mean Squared Error (RMSE), and correlations with true θ values, using the GRM. Ho (2010) also found MFI and Bayesian methods performed comparably in terms of mean bias, RMSE, absolute average difference, and correlations with true θ values, using the GPCM and a number of θ estimation methods.

Han (2009) recently proposed gradual maximum information ratio (GMIR) as the latest alternative item selection method to MFI. A few studies have compared

GMIR to MFI under the three-parameter logistic model (Han, 2009; Han, 2010) and under the Generalized Partial Credit Model (Chang & Dodd, 2013). Han (2009) found GMIR resulted in a smaller standard error of θ estimate on average over 20 days of administration. Han (2010) found that GMIR, MFI, and KL outperformed IIC and LWIC especially with a shorter test, 10-20 items. Chang and Dodd (2013) found that GMIR had better measurement precision than MFI at extreme trait values and in the early stages of the CAT. Differences were found in overall measurement precision only at extreme θ values. If patients with the illness of interest have extreme trait values, poor shoulder functioning for example, greater measurement precision at those values, would be extremely important. This would ensure the patients' functioning levels are measured precisely. Chang and Dodd also found the differences between MFI and GMIR were present for the first 5-10 of the 20-item measure. This increase in measurement precision in the first few items could be especially advantageous in PRO measures, since very few items are administered. GMIR could allow fewer items to be administered with greater measurement precision. This dissertation research aims to extend the research of GMIR comparing it to MFI in a CAT simulation under ARSM and conditions similar to health care PRO measures: 5-10 items administered and matched and mismatched population distributions. This study will use item parameter estimates from an operational PRO measure item bank and simulated response data.

Chapter 2: Literature Review

The first section presents item response theory (IRT), it's advantages and assumptions, and models appropriate for dichotomous and polytomous items. The second section is a discussion of computer adaptive testing (CAT), its advantages, and the major components of a CAT. The third section contains a review of CAT item selection method research. The fourth section is the statement of problem and research questions for the dissertation.

ITEM RESPONSE THEORY

Item response theory (IRT) is a model-based measurement approach that uses examinee's item responses and item properties to estimate latent traits, e.g. ability, attitudes or functioning level. IRT is also called latent trait theory. Each model includes a person parameter, which is the person's latent trait score (θ), and one or more item parameters. Depending on the restrictiveness of the assumptions of the model, item parameters can include the item difficulty, item discrimination, and a pseudo-guessing parameter. The model, with these parameters included, predicts the probability of a particular response to an item. The probability of a response is a function of the person and item parameters, as defined by the model. IRT models with binary responses, e.g. right or wrong, are called dichotomous models and models with multiple-category responses, e.g. 1,2,3,4,5, are called polytomous models.

One advantageous feature of IRT is parameter invariance. Regardless of the items administered, the person's trait measurements will vary only within a linear transformation and regardless of the trait level of the people measured, the item parameters will vary only within a linear transformation. In other words, a person's score would be the same across different sets of items administered, and item calibrations would be the same across different populations of test takers, accounting for sampling error.

Dichotomous IRT Models

Dichotomous IRT models predict the probability of a response to an item, conditional on the person and item parameters, when the possible response options are binary (e.g. right or wrong, 0 or 1). Frequently used, dichotomous IRT models include the one-parameter-logistic model, two-parameter logistic model, and three-parameter logistic model, all of which are named for the number of parameters within each model. Each model is described below.

One-parameter logistic model

The one-parameter logistic model (1PL) or Rasch (1960) model is the simplest model. It predicts the probability of a correct response based on only the person's trait estimate and the item difficulty parameter. The equation for the 1PL is

$$P(u_i = 1|\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \tag{1}$$

Where u_i is the examinee's response to the item, θ is the examinee's trait estimate, and b_i is the difficulty parameter of item *i*. The item difficulty is defined as the θ value at the point of inflection of the item characteristic curve, which is always .5 for the 1PL model. Hence, when the θ matches the item difficulty, the examinee has a .5 probability of answering the item correctly. The item difficulty parameters and trait estimates are on the same scale, generally ranging from -3 to +3.

Two-parameter logistic model

Birnbaum's (1968) two-parameter logistic model (2PL) includes two item parameters: a difficulty parameter like the 1PL model as well as a discrimination parameter. The model predicts the probability of a correct response based on the examinee's trait estimate, the item difficulty parameter, and the item discrimination parameter. The equation for the 2PL is

$$P(u_i = 1|\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{2}$$

Where u_i is the examinee's response to the item, θ is the examinee's trait estimate, D is the scaling constant 1.7, a_i is the discrimination parameter of the item *i*, and b_i is the difficulty parameter of item *i*. A larger item discrimination parameter indicates the item's greater ability to discriminate between examinee's trait estimate levels or stronger relationship to an examinee's level of a latent trait. It is also proportional to the slope of the item characteristic curve at the point of inflection. The discrimination parameter generally ranges from 0 to 2 in practice. When a_i=1, the 2PL model simplifies to the 1PL model.

Three-parameter logistic model

Birnbaum's three-parameter logistic model (3PL) includes the item difficulty and discrimination parameters of the 2PL model and adds a pseudo-guessing or pseudo-chance parameter. The model predicts the probability of a correct response based on the person's trait estimate, the item difficulty and discrimination parameters, and the pseudo-chance parameter. The equation for the 3PL is

$$P(u_i = 1|\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{3}$$

Where u_i is the examinee's response to the item, θ is the examinee's trait estimate, c_i is pseudo-chance level of item *i*, the D is the scaling constant 1.7, a_i is the discrimination parameter of the item *i*, and b_i is the difficulty parameter of item *i*. The pseudo-chance parameter is the lower asymptote of the item characteristic curve, or the probability of an examinee with a very low trait level selecting the correct response to a very difficult item. When c_i=0, the 3PL model simplifies to the 2PL model.

Item Characteristic Curve.

Item characteristic curves mathematically illustrate the relationship between probability of a correct response and the person's trait estimate, or θ. The probabilities are plotted against the θ values to produce the s-shaped item characteristic curve. Because the 1PL, 2PL, and 3PL models differ in terms of parameters, the item characteristic curves also differ; *Figure 1* below shows typical ICCS for each model.

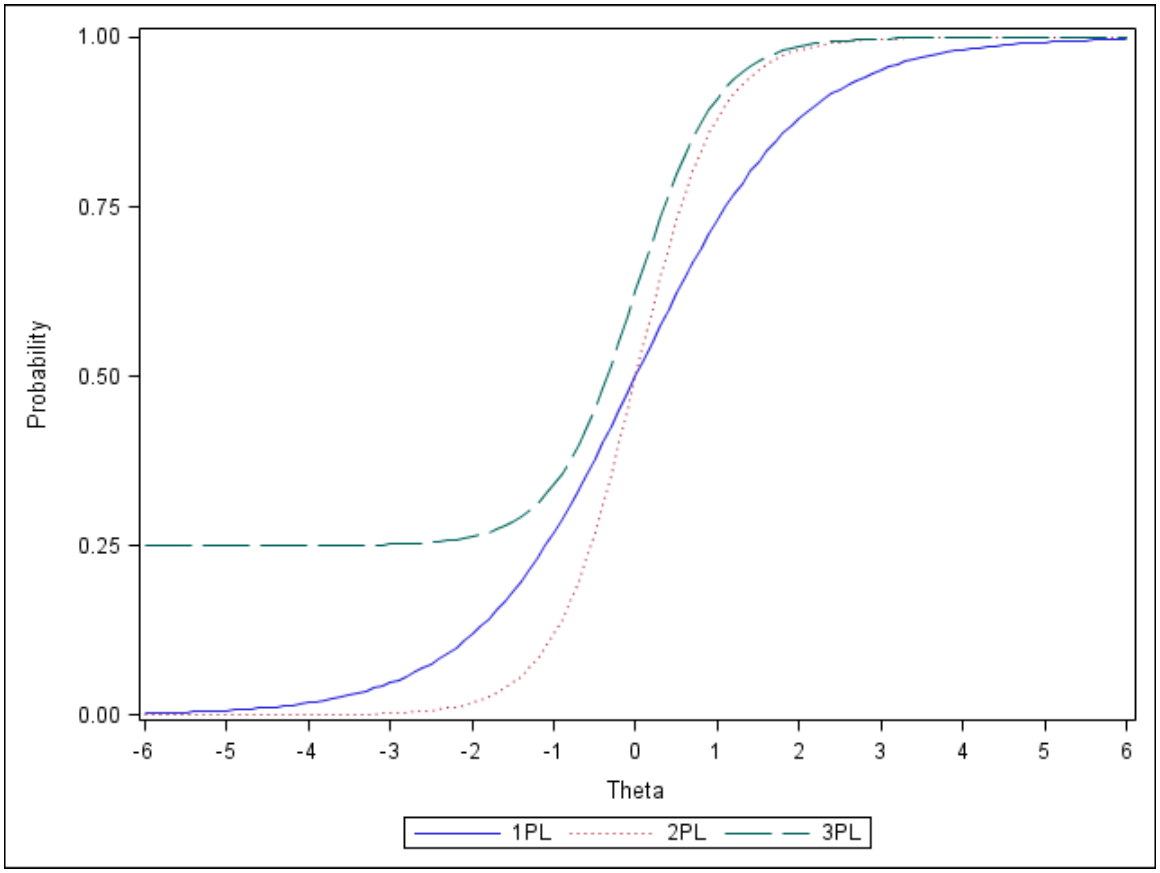


Figure 1. ICCs for 1PL, 2PL, and 3PL Models
Note: For the 1PL, b=0.0. For the 2PL, b=0.0 and a=2.0.
For the 3PL, b=0.0, a=2.0, and c=0.25.

ICCs for 1PL vary by the item difficulty parameter or the location of the curve along the θ scale. The b_i value equals the θ value at the point of inflection of the ICC, when the probability of a correct response is .5. Items to the right on the θ scale are more difficult than items to the left. 1PL ICCs have identical slopes and lower asymptotes at 0.0.

The ICCs for the 2PL vary by the item difficulty parameter, like the 1PL ICCs, but the 2PL ICCS also vary by the item discrimination parameter. As with the 1PL, the b_i value equals the θ value at the point of inflection of the ICC and the lower asymptote is at 0.0. The item discrimination parameter, a_i, represents the difference of items in discrimination. The slope for the item is 0.425 a_i.

Like the 2PL ICCs, the ICCs for the 3PL vary by the item difficulty and discrimination parameters, however the 3PL ICCs also vary by a pseudo-chance parameter. The pseudo-chance parameter, c_i, is the lower asymptote for the ICC. Since the lower asymptote is no longer 0.0 as with the 1PL and 2PL, the point of inflection is no longer .5. It is now $\frac{1+ci}{2}$ or half the distance between c_i and 1. The b_i value still equals the θ value at the point of inflection of the ICC.

Polytomous IRT Models

Polytomous IRT models predict the probability of a particular response to an item, conditional on the person and item parameters, when there are more than two possible response options (e.g. 0,1,2,3,4,5). These models could be used for example for a Likert-type attitude scale or partial-credit scoring of a test item. Polytomous models extend from dichotomous models; instead of a single item difficulty parameter, polytomous models use multiple item step difficulty, category boundary or item threshold parameters. Polytomous IRT models can be separated into two classifications: difference models and divide-by-total models (Thissen & Steinberg, 1986).

Difference models include the Muraki's (1990) Rating Scale Model (MRSM) and the Samejima's (1969) Graded Response Model (GRM). The GRM is appropriate for essays and partial credit scoring. It includes a category boundary parameter and an item discrimination parameter. The MRSM is a restricted case of the GRM and is appropriate for attitude measurement. The MRSM expands the category boundary parameter of the GRM into a location parameter and a set of threshold parameters for the entire scale. If there are only 2 response categories, both the GRM and MRSM reduce to the 2PL model. For difference models, the conditional probability of a scoring in a particular response category *x,* or $P_{ix}(\theta)$, is calculated in two steps. First, calculate P^*_{ix} (θ), the conditional probability of responding in category *x* or higher on item *i*. This is generally an exponential divided by one plus the exponential. The second step is to calculate $P_{ix}(\theta)$ by subtracting P^*_{ix} (θ) for adjacent categories. For example, with three response categories (*x*=0,1,2), the probability of responding in category 1 is P_{i1} (θ)= P^*_{i1} (θ)- P^*_{i2} (θ), where $P^*_{i1}(\theta)$ is the probability of responding in category 1 or 2 and P^*_{i2} (θ) is the probability of responding in category 2. The probability of responding in or above the lowest category is $P^*_{i0}(\theta)=1$, and the probability of responding above the highest category is $P^*_{i3}(\theta)=0$ by definition.

Divide-by-total models include Andrich's (1978) Rating Scale Model (ARSM), Rost's (1988) Successive Intervals Model (SIM), Master's (1982) Partial Credit Model (PCM), Muraki's (1992) Generalized Partial Credit Model (GPCM), and Bock's (1972) Nominal Response Model (NRM). In divide-by-total models, the conditional probability of scoring in a particular response category *x* equals an exponential divided by the sum of the exponentials for conditional probabilities for all categories.

Under the PCM, the conditional probability of scoring in a category includes only a set of item step difficulty parameters. The PCM assumes all items are equally discriminating. Therefore, the PCM simplifies to the 1PL or Rasch model when there are only two item responses. The GPCM expands the PCM by removing the equal discrimination assumption. So, when the discrimination parameter is 1.0, the GPCM simplifies to the PCM. When there are only two response alternatives, the GPCM simplifies to the 2PL model.

ARSM and SIM are both appropriate for attitude measurement. Both the ARSM and the SIM are simplifications of the PCM. The SIM includes a dispersion parameter, allowing the thresholds to vary among items across the scale. When this dispersion parameter is 0.0, the SIM simplifies to the ARSM. Since most attitude scale thresholds do not vary, the ARSM is generally sufficient. The NRM is the most general divide-by-total model and is generally used with multiple-choice items, when the detractors aren't ordered in degree of correctness. When the slope parameter of the NRM is constrained, the PCM is obtained (Thissen & Steinberg 1986). When there are only two response categories, the NRM simplifies to the 2PL. The PCM and the ARSM, the models most relevant to this study, are described in more detail below.

Partial Credit Model

Master's (1982) PCM is appropriate for essay and partial-credit scoring. The PCM assumes there is not guessing and no items differ in discrimination. While, the PCM requires the item steps to be completed in order, but the steps do not need to be in order along the θ scale, from low to high or easiest to hardest. The equation for the conditional probability of responding in category *x* for item *i* is

$$P_{ix}(\theta) = \frac{\exp[\sum_{k=0}^{x}(\theta - b_{ik})]}{\sum_{h=0}^{m_i}\exp[\sum_{k=0}^{h}(\theta - b_{ik})]} \qquad (4)$$

where b_{ik} is the item step difficulty parameter associated with the transition from one category to the next and m_i is the number of step difficulties for item *i*. There are m_i +1 possible score responses for item *i*, (i.e., x = 0, 1, . . . , m_i).

Andrich Rating Scale Model

Andrich's (1978) Rating Scale Model is appropriate for items that share a common rating scale, e.g. Likert-type scale items. It is a simplification of the PCM. The item difficulty parameter can be decomposed into the location for each item on the latent trait scale and one set of threshold parameters dividing categories for the entire set of items in the scale. So, each item varies in relative easiness or difficulty, but the set of thresholds between categories are constant for the entire scale. Using a Likert-type scale example having response options 1= disagree, 2= neutral, and 3=agree, the relative locations of these response categories would be expected to remain constant across an entire scale. The ARSM assumes all items to have equal discrimination and does not include a pseudo-chance parameter. The equation for the conditional probability of responding in category *x* for item *i* is

$$P_{ix}(\theta) = \frac{e^{[K_x + x(\theta - b_i)]}}{\sum_{h=0}^{m_i} e^{[K_h + h(\theta - b_i)]}} \qquad (5)$$

Where K_x is the negative sum of thresholds passed, *x* is the category score, θ is the examinee's trait estimate, b_i is the difficulty parameter of item *i, m_i* is the number of categories, *h* is every category, and K_0=0. Since the ARSM is a Rasch model, assuming equal discrimination and not including pseudo-chance, the raw score is a sufficient statistic to estimate an examinee's trait.

Category Response Curves

Category Response Curves (CRC) illustrate the probability of an examinee responding in a particular category, conditional on θ. The probabilities of responding in the first and last categories are monotonic functions and the middle category or categories are non-monotonic symmetric functions. *Figure 2* below displays CRCs for a polytomous item.

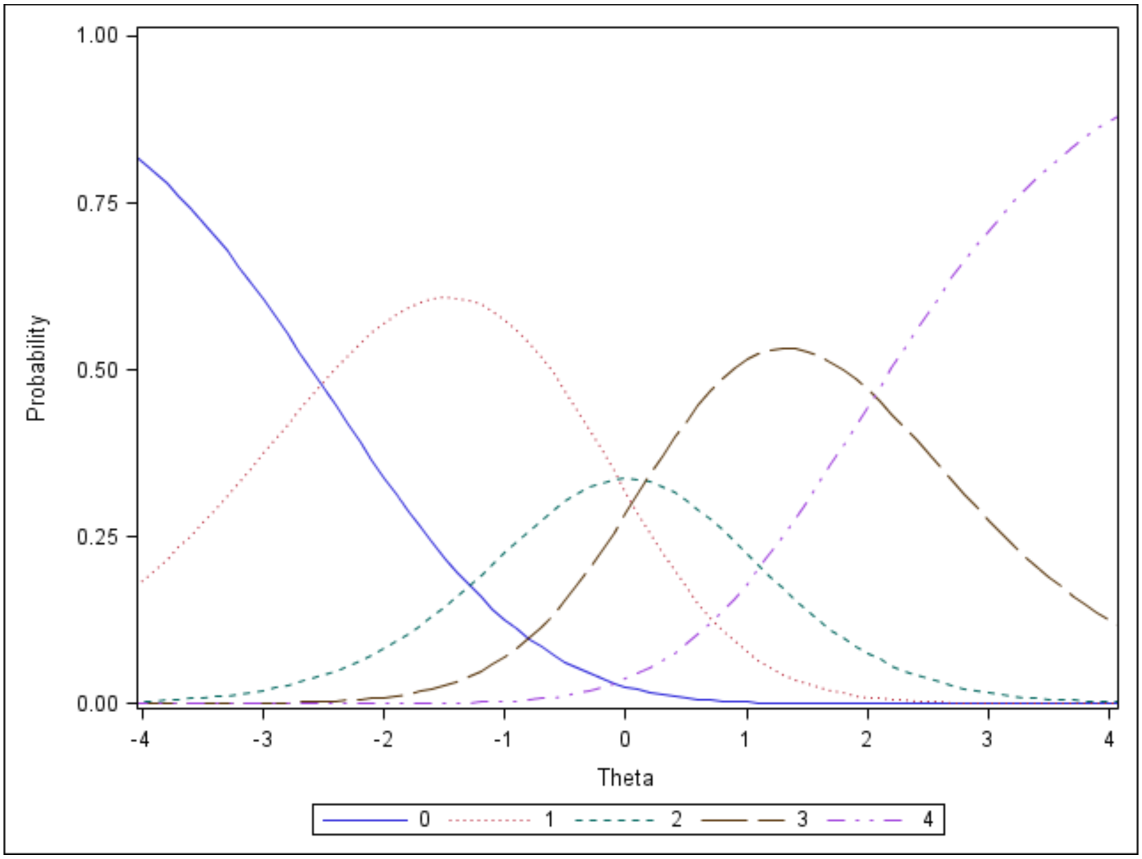


Figure 2. CRCs for a Five-category Item
Note: 0,1,2, 3, and 4 are the category scores.

The shape and location of the curves are determined by the item parameters. In general, a higher discrimination parameter results in more narrow and peaked

CRCs and means higher differentiation among trait levels. The response probabilities sum across categories to 1.0 for any fixed θ value.

Assumptions

Unidimensional dichotomous and polytomous IRT models all involve three assumptions: unidimensionality, local independence, and functional form. A test that is unidimensional measures a single trait, or θ. If a test measures multiple traits or constructs, this assumption is violated and a multidimensional IRT model would be more appropriate. For more information about multidimensional IRT models, see Reckase (2009). Local independence is achieved when the probability of a response on any one item is independent of the outcome of any other item, controlling for θ and the item parameters. Specifically, strong local independence means the item responses are statistically independent, conditional on θ. Weak local independence means the item responses are uncorrelated, conditional on θ. Item responses should at least meet weak local independence. Local independence is violated if a response to one item is related to other items; this could overestimate the precision of measurement. For example, if an item gives information toward another item's response or items are linked by a common prompt, e.g. a reading passage; those items would not be independent. If items are grouped according to a common stimulus and dichotomously scored, the testlet response theory (TRT) model is appropriate(Wainer, Bradlow, & Wang, 2007). To meet the functional form assumption, item characteristic curves, for dichotomous models, and category characteristic curves, for polytomous models, mathematically depict the relationship between probability of a response and the person's trait estimate, or θ.

The curves illustrate how a change in an examinee's trait estimate relates to a change in the probability of a particular response. These curves vary only as a function of the model's specified item parameters, not across examinees.

ITEM AND TEST INFORMATION

An Item Information Curve (IIC) indicates the amount of Fisher information a dichotomous or polytomous item provides across the latent-trait scale. The equation for the amount of Fisher information provided by a dichotomous item at θ is

$$I_i(\theta) = \frac{P'_i(\theta)^2}{P_i(\theta)(1 - P_i(\theta))} \tag{6}$$

where $P_i(\theta)$ is the conditional probability of responding correctly to item *i* and $P'_i(\theta)$ is the first derivative of $P_i(\theta)$ (Birnbaum, 1968). The equation for the amount of Fisher information provided by a polytomous item at θ is

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{(P'_{ix}(\theta))^2}{P_{ix}(\theta)} \tag{7}$$

where $P_{ix}(\theta)$ is the conditional probability of responding in category *x* to item i and $P'_{ix}(\theta)$ is the first derivative of $P_{ix}(\theta)$. The information denotes how effectively the item measures the latent trait. The amount of information provided by an item is maximized around the item-difficulty parameter for the 1PL and 2PL and slightly below the item-difficulty parameter for the 3PL. If the item's difficulty matches the examinee's ability on the θ scale, the item information is greater. Additionally, for dichotomous items, the higher the item-discrimination parameter, the greater the information provided at that θ. The IIC with a higher item-discrimination parameter will appear more 'peaked', or higher and narrower than a lower item-discrimination parameter, which would appear flatter and wider. For polytomous items when using the PCM, items with a greater number of and larger magnitude of reversals in step difficulties will appear more peaked (Dodd & Koch, 1987).

If items are calibrated on a common latent-trait scale, they can be added across items within a test to calculate the test information. Test information is therefore the sum of item information functions. The equation for test information is

$$TI(\theta) = \sum_{i=1}^{n} I_I(\theta) \tag{8}$$

where n is the number of items and I(θ) is the information provided by item i at a given θ. The test information determines how effectively a set of items is measuring a latent-trait. The test most precisely measures the θ at the peak of the test information curve. The test information has a direct relationship with the examinee's standard error of measurement. The equation for the standard error of measurement at a given θ estimate is

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \tag{9}$$

The test information and standard error of measurement are inversely related; for example, when the test information is high for a given θ, the standard error of measurement is low.

COMPUTER ADAPTIVE TESTING

Computer Adaptive Testing (CAT) is a form of computer-based testing that adapts to the examinee's trait level. In contrast to traditional testing, which administers an identical set of items as a particular form of the test to every examinee, the CAT program selects and administers items that match the examinee's currently estimated trait levels, essentially administering many forms of the test.

Advantages of CAT

CAT has a number of advantages over traditional testing. The efficiency of the test is increased by the adaptive process. As compared to traditional P&P tests, CATs can be shorter in length while maintaining comparable or greater measurement precision. (Dodd et al., 1995; Embretson & Reise, 2000; Fliege et al., 2005; Wainer, 2000; Ware et al., 2003; Ware et al., 2005). Examinees don't waste time and energy on items that are not appropriate, unlike tradition P&P testing that administers items that match the examinees trait level as well as items that are a mismatch to the examinee. For example, commonly-used non-adaptive, versions of shoulder functioning scales contain only 8-20 items (Cook, Gartsman, Roddey, & Olson, 2001). Items that are not a match to the examinees trait level, e.g. items that are too difficult for a low functioning patient, do not provide any information about the patient. Ware et al. (2003) found a 6 item CAT (CAT-HIT) correlated highly with a 54-item total item Headache Impact Test, while allowing a 90.8% reduction in response burden. The CAT-HIT did not have 'ceiling' or 'floor' effects, had high reliability estimates and clinical validity. The CAT-HIT was also more responsive at monitoring changes in headache impact over time. Fries et al. (2014) also found reduced ceiling and floor effects: the CAT covered 6.3 SD at a reliability of 0.90 or better as opposed to only 2.4 and 4.8 for the classical measures and static IRT measure respectively. Ware et al. (2005) evaluated the use of CAT to measure rehabilitation outcomes and concluded it could improve the measurement of physical functioning in rehab settings. They also found patients preferred the self-administered CAT to an interviewer-administered, static survey. Similarly, Koch, Dodd, and Fitzpatrick (1990) found the majority of students preferred a CAT administration and though it was more fun than the paper and pencil version of a survey about alcohol attitudes. The majority also felt the computer survey method would result in more honest answers.

Another advantage of CAT is that the number of patients needed for a clinical trial can be reduced by using the most informative items in a CAT instead of an identical set of items to every subject, while maintaining statistical power (Fries, 2006; Fries et al., 2014). Fries and colleagues compared a 10 item CAT to a 20 item static short form (derived using IRT methods) and two "legacy" health measures, the Health Assessment Questionnaire (HAQ) and the 10 item physical function instrument (PF-10), which were developed and evaluated under classical test theory. They found the CAT only required 100 subjects in order for statistical significance at the $p<0.05$ level as opposed to 427 subjects. The reduction in the number of subjects, allows for a reduction in resources and funding necessary for research trials. Similarly, in a research setting and in practices, item, measures, and evaluation results are easier than P&P to integrate into patients' electronic records and to share across institutions (Walker et al., 2010). Fewer resources are required for better items, measures, and patient records to be stored and shared across settings. Measures and results can more easily be compared across settings.

Item pool, item selection method, trait estimation, and stopping rule are the four major components to a CAT, regardless of the IRT model used (Reckase, 1989; Wainer, 2000).

Item Pool

A CAT item pool consists of the entire bank of items that were calibrated using an IRT model and could be administered to an examinee. The size and psychometric characteristics of the item pool influence the properties of the CAT. The necessary size of the item pool, or number of items available to administer, depends on the item response type (dichotomous or polytomous) and inclusion of content balancing and exposure control procedures. CATs with dichotomously scored items require a larger number of items in the pool than polytomously scored items. Polytomous item pools can be smaller because each item provides more information across a wider range of the trait scale (Dodd et al., 1995). Each pair of adjacent categories of a polytomous item is equivalent to one dichotomous item. While dichotomously scored item pools may need several hundred to over a thousand items, polytomously scored item pools can be as small as 30 (Dodd, 1990; Dodd & de Ayala, 1994). Van Rijn, Eggen, Hemker, and Sanders (2002) using the GPCM to compare the maximum Fisher information and interval information criteria methods found that a 500 item bank had resulted in a smaller RMSE of ability estimates than the 150 item bank, but there was not an interaction between the item bank size and the item selection methods. Lima Passos, Berger, and Tan (2007), using the nominal response model, found item pool size and quality affected the stability of the ability estimates in the first 5-10 items, especially at the extremes. Item pools with 600 items had reduced magnitude in fluctuations of RMSE and BIAS compared to the 300 item pools, but the efficiency after all 15 items did not differ between the two pool sizes. Additionally, the flat shaped item pools had less magnitude in fluctuations than the bell shaped item pools mainly at the extreme trait values. Item pool size in studies of PRO measures varies. Ware et al. (2005) used a 77 item pool for a rehabilitation outcome study. Fliege et al. (2005) ultimately used a 64 item pool developing a CAT for depression. Ware, Bjorner, and Kosinski (2000) have 53 items in their item pool for the Headache Impact Test (HIT). In a review of IRT-based CAT development for PRO measures studies, Walker et al. (2010) found item banks contained between 12 and 282 items, with a median of 50 items.

In high-stakes testing, where content validity and test security are important, content balancing and item exposure control procedures can be used. Item pool sizes under these constraints should be larger than unconstrained CAT item pools. Davey and Nering (2002) state that CAT item pools should contain the equivalent of 5-10 conventional forms at a minimum. Studies have found that at least 100 to 120 polytomous items are necessary to achieve the desired exposure rates and avoid overexposure (Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; LeRoux, Lopez, Hembry, & Dodd, 2013; LeRoux & Dodd, 2014; McClarty, Sperling, & Dodd, 2006). Additionally, if specific content areas contain very few items, more items may be necessary to avoid overexposure of those items (Burt, Kim, Davis, & Dodd, 2003; Davis & Dodd, 2003).

In addition to the size of the pool, psychometric properties of the item pool influence the properties of the CAT. The item pool information function is the sum of the information functions for all the individual items in the pool. Commonly, CAT item pools provide information for the average individual in the population, with the most information in the middle of the trait scale. However, subpopulations might have trait levels at the extremes of the scale. In specific assessment purposes with targeted subpopulations, an item pool targeting the average individual in the population could not be ideal. For example, testing in an educational setting evaluates students with low abilities for remediation or high abilities for talented and gifted programs. Similarly, examinees being tested for a medical outcome may not align with the trait levels of the general population, or a post-treatment assessment subpopulation might not match with the pretreatment assessment values. Instead of pools that provide maximum information in limited ranges of θ, ideally, the item pool would provide information across the entire θ scale at the extremes in addition to the middle. Some previous research found that a match between the population distribution and the item pool distribution results in more accurate θ estimates using MFI, but others found this impact is minimal (Chang & Dodd, 2013; Dodd, Koch, & de Ayala, 1993; Gorin, Dodd, Fitzpatrick, & Shieh, 2005; Keng, 2008; Lee & Dodd, 2012). Lee and Dodd (2012) found that measurement precision of the polytomous CAT using PCM was relatively robust to a mismatch between item pool distributions and trait distributions.   Chang and Dodd (2013) found that whether the population distribution was normal or negatively skewed, when the item pool distribution was relatively normal, did not interact with the item selection method.

Other studies have shown that trait estimates are more accurate when the item pool information distribution aligns with the trait distribution of the examinees (Dodd et al., 1993; Gorin et al., 2005; Keng, 2008). Dodd et al. (1993) found a match between the distribution of examinee ability and the item pool resulted in more accurate trait estimates based on the PCM with MLE. Also, nonconvergent cases occurred more when there was a mismatch and MLE was used. With a normally distributed participant population, larger numbers of nonconvergent cases occurred for easy and hard item pools, as opposed to items that spanned the continuum or included half easy and half hard items. Easy and hard items pools had information functions peaked at the low end and high end of the trait dimension, respectively. Item pools spanning the continuum and half easy, half hard item pools had information functions peaked at the middle and bimodal, respectively.

Gorin et al. (2005) also found lower average standard errors, fewer items administered on average, and higher correlations between known and estimated θ values when the item pool information and the examinee trait distribution aligned and the item pool covered the entire θ range using the PCM. The item pool that covered the entire range was compared to pools with only easy or hard items. This was true for normal and skewed examinee trait distributions. Gorin et al. (2005) found issues with nonconvergence with maximum likelihood estimation (MLE) and weighted likelihood estimation (WLE) when the items did not cover the full θ scale. While MLE, WLE, and expected a posteriori (EAP) all performed poorly with a mismatched item pool for skewed distributed trait population, EAP performed well for a normally distributed trait population with a misaligned item pool. However, Keng (2008), investigating multi-stage testing and testlets in a constrained CAT, also found that reducing the item pool from 46 to 31 testlets (1,008 and 741 items respectively) did not affect the measurement precision, but he found that reducing the item pool while also creating a mismatch between population ability and item pool distribution resulted in increased bias and RMSE. This increase was more pronounced when, additionally, the test length was reduced from 42 to 21 items.

Item Selection

CAT items can be selected based on their psychometric properties, content balancing, and/or exposure control. Depending on the purpose of the test and the proposed use of the scores, content balancing and exposure control may or may not be necessary. Content balancing ensures that the content test specifications are met, the correct proportion of items for different content areas. The most commonly used content balancing procedure is the Kingsbury and Zara (1989), which compares target content proportions from test specifications to the actual proportions during the test administration. The next item administered comes from the content area with the largest discrepancy between target and actual proportion. When test security is an issue, exposure control ensures items are not overexposed by being administered to a very large number of examinees (Boyd, Dodd, & Choi, 2010).

Exposure control methods fall into four types: randomization, conditional, stratified and combination. Randomization procedures select a group of a certain number of maximally-informative or near maximally-informative items and randomly select an item from that group to administer. Randomesque (Kingsbury & Zara, 1989) and modified-within-.10-logits procedure (Davis & Dodd, 2003) are randomization procedures. Sympson-Hetter (Sympson-Hetter, 1985) and conditional Sympson-Hetter (Stocking & Lewis, 1998) are conditional procedures, involving simulations to set exposure control parameters for each item to limit exposure of the most likely to be administered items below the maximum exposure rate. Stratified procedures divide items into strata based on the discrimination parameter and additionally on the difficulty parameter, the a-stratified procedure (Chang & Ying, 1999) and the a-stratified with b blocking procedure respectively. Items with higher discrimination parameters are saved for later in the test when the trait estimate is more accurate. Combined procedures are methods that combine multiple procedures. For example, the progressive-restricted (Revuelta & Ponsoda, 1998) selects items on both information and a random component, relying more on information as the test progresses, and restricts items if the maximum exposure rate is surpassed. A second combined procedure, the enhanced stratified method (Leung, Chang, & Hau, 2002) within a-stratified strata sets the Sympson-Hetter exposure control parameters though simulations. When content balancing and or exposure control are used in item selection, the CAT is a constrained CAT.

If items are selected only on the psychometric properties of the items, the CAT is referred to as an unconstrained CAT. In the medical field, patient reported outcome measure CATs can select items solely on psychometric characteristics; exposure control and content balancing procedures are not necessary. One example of a psychometric property of an item is the information an item provides at an examinee's given trait level. A variety of item selection methods have been developed. Maximum Fisher information is the most commonly used method. Item selection methods are discussed in more detail below.

Item Selection Methods

There are two general approaches to item selection methods: information-based and Bayesian (van Rijn et al., 2002). Information-based item selection methods chose the item that is most informative at a specific θ estimate, based on Fisher information or Kullback-Leibler information. Bayesian item selection methods chose the item based on the prior and posterior distributions of θ estimates. All item selection methods were originally developed and studied with dichotomous items, but have been applied to polytomous items recently (Choi & Swartz, 2009; Ho & Dodd, 2012; Penfield, 2006; van Rijn et al., 2002; Veldkamp, 2003). In this section, this proposal will describe information-based item selection methods: maximum Fisher information, general weighted information, Kullback-Leibler information, and gradual maximum information ratio. It will also describe Bayesian item selection methods: Owen's approximate Bayesian, maximum posterior weighted information, maximum expected information, minimum expected posterior variance, and maximum expected posterior weighted information.

Maximum Fisher Information

The most commonly used item selection method for dichotomous and polytomous CATs is maximum Fisher information (MFI) criterion (Lord, 1980). As discussed in the IRT section, Fisher item information $I_i(\theta)$ represents how effectively or precisely the item measures the examinee's trait level, for a given θ. MFI selects the item that maximizes the Fisher's information at a given trait estimate. While the MFI method is easy to employ, the main weakness of the method is the attenuation paradox (Lord & Novick, 1968); the MFI depends on a match between the current θ and the true θ of the examinee. A mismatch leads to suboptimal item selection; the items selected are the most informative items for the examinee's interim θ, but not necessarily the examinee's true θ. This leads to inefficient θ estimates. So, the CAT measurement precision is compromised. This is especially problematic at the early stages of a CAT, when only a few items have been administered or for a CAT with very few items. The interim θ estimate at this point might be far from the true θ. To overcome this weakness, alternative item selection methods have been developed. These alternatives are discussed below.

General Weighted Information

To address the attenuation paradox, Veerkamp and Berger (1997) proposed the general weighted information (GWI) criterion. The criterion is formulated as the weighted average of the information function values over all possible θ values:

$$GWI_i(\theta) = \int_{-\infty}^{\infty} W(x_n|\theta)I_i(\theta)d\theta \tag{10}$$

$$GWI_i(\theta) = \sum_{j=1}^{k} W_n(x_n|\,\theta_j)I_i(\theta) \tag{11}$$

Equation 9 is for a continuous θ scale and equation 10 is for a discrete θ scale. $W(x_n|\theta)$ is a weighted function, where x_n is the vector of responses. $I_i(\theta)$ is the Fisher information for item i at a given θ. In equation 9, integral $\int_{-\infty}^{\infty} d\theta$ indicates the area over all possible θ values and in equation 10 k is the quadrature points. By incorporating the information over a range of θ values, the GWI incorporates the uncertainty of the interim θ estimate using a specific weight function. The GWI criteria select the item that maximizes the GWI.

Veerkamp and Berger (1997) developed two variations of the GWI by varying the $W(\theta)$, interval information criterion (IIC) and the likelihood weighted information criterion (LWIC). IIC uses a confidence interval around the true θ estimate, weighting each θ level in that confidence interval uniformly and values outside the confidence interval as 0. The IIC is formulated as the area under the Fisher information function from $\hat{\theta}_L$ to $\hat{\theta}_R$, or,

$$\int_{\theta=\hat{\theta}_L}^{\hat{\theta}_R} I_i(\theta)\, d\theta \tag{12}$$

where $\hat{\theta}_L$ and $\hat{\theta}_R$ are the left and right limits of the confidence interval, respectively. The LWIC gives more weight to the information function when the interim θ estimate is close to the true θ. The LWIC is the area that is the product of the Fisher information and the likelihood function, or

$$\int_{-\infty}^{\infty} L_n(\theta|x_n) I_i(\theta) d\theta \tag{13}$$

where $L_n(\theta|X_n)$ is the likelihood function after administering n items with the response pattern x_n.

In dichotomous studies, Veerkamp and Berger (1997) and Chen, Ankenmann, and Chang (2000) found the interval information criterion (IIC) θ estimates had poorer measurement precision on short tests. The IIC had larger bias, mean squared error (MSE), root mean squared error (RMSE), and standard error (SE) at the early stage of the CAT (1-10 items), especially at the extreme θ values, under a 3PL model. These differences decreased as the test length increased; after the first 10 items, the performance was comparable. Han (2010) also found poorer measurement precision by the IIC; it had inconsistent standard error of the θ estimate (SEE) and larger mean absolute error (MAE) in the middle of the θ values. Han did find better item pool usage by the IIC than MFI. Chen et al. (2000) found that MFI and IIC did not select the same initial item for administration (item overlap), but proportion of item overlap increased from .00 on the first item to .50 on a 5 item test and ultimately to a .80 on a 20 item test, on average across θ values. Item overlap was generally higher for higher θ values, especially on shorter tests.

In polytomous studies, while IIC did not have poorer measurement precision, it was not an improvement on MFI (van Rijn et al., 2002; Veldkamp, 2003). Under the GPCM, these studies found comparable bias, MSE, and RMSE of θ estimates on 10 item and 30 item tests. Using the GPCM, Veldkamp (2003) found high item overlap, between 85-100%, and the small differences diminished as the item pool size and average discrimination decreased. Lima Passos et al. (2007) found, under the NRM, the RMSE and Bias for the IIC fluctuated more than the Kullback-Leibler information and MFI for the first 5-10 items. This instability was especially at extreme θ values and using the bell-shaped and smaller item pools as opposed to flat and larger pools. While the IIC was less stable and less efficient at the first few items, the results converged after 10-15 items.

Veerkamp and Berger's (1997) second general weighted information approach, likelihood weighted information (LWI), has shown mixed results. In dichotomous studies, Veerkamp and Berger found the LWI outperformed the MFI at the early stage of the CAT at the extreme θ values using MLE estimation under a 3PL model. Specifically, LWI had smaller bias and MSE values than MFI. Differences decreased as the test length increased. Using EAP estimation, LWI performance was comparable to MFI. Han (2010) found the LWI had larger SEE at the negative extreme θ values and larger MAE at both extremes, using MAP. However, Han found the LWI had better level of item pool usage than the MFI across all test lengths.

In a polytomous study, Choi and Swartz (2009) found LWI performed comparably to MFI and other approaches, using EAP θ estimation and the GRM. The item selection methods had similar overall bias, RMSE, and correlations with true θ values; especially as the test length increased from 5 to 10 to 20 items. Even when only 5 items were administered, there was not an item selection method that had clearly better performance.

Kullback-Leibler Information

As an alternative to Fisher information, Chang and Ying (1996) proposed the Kullback-Leibler (KL) information. KL information for item *i* can be expressed as

$$K_i(\theta \parallel \theta_0) = P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + [1 - P_i(\theta_0)]\log\left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)}\right] \qquad (14)$$

where θ $_0$ is the true θ and the ‖ signifies that θ must be different from θ $_0$. The KL information value represents how well an item differentiates between these two θ levels. In contrast to the Fisher information, which represents the precision of measurement at the examinee's estimated θ, a single value, KL information represents the precision across a range of θ values. Chang and Ying (1996) noted that the KL function is not symmetric, KLj (θ ‖ θ $_0$) ≠ KLj (θ $_0$‖ θ). Also, when the two θ values are equal, the KL information is 0; when they are different, the KL information is greater than 0.

For item selection, the KL information criterion chooses the item with the maximum area under the KL function within the confidence interval bound by $\hat{\theta}_L$ and $\hat{\theta}_R$. The KL function is expressed by

$$\int_{\hat{\theta}_L}^{\hat{\theta}_R} K_i(\theta \parallel \hat{\theta}_n)\, d\theta \tag{15}$$

where $\hat{\theta}_n$ is the estimated θ after *n* items. The criterion transitions from KL information to Fisher information by the confidence interval decreasing, as the number of items administered increases. Chang and Ying (1996) also proposed a Bayesian item selection method that weights the KL information by the posterior θ distribution. The Kullback-Leibler information with a posterior distribution (KLP) is expressed by

$$\int_{-\infty}^{\infty} g(\theta|x_n) K_i(\theta \parallel \hat{\theta}_n) d\theta \tag{16}$$

where g(θ|x_n) is the posterior distribution after administering *n* items with the response pattern x_n.

In dichotomous CATs, criterion using Kullback-Leibler information, KL and KLP, performed better than MFI in the early stages at extreme θ values specifically. (Chang & Yang, 1996; Chen et al., 2000). Chen et al. (2000) found KL and KLP had smaller bias and SEs at extreme negative θ values with 5 or fewer items. Bias and SEs were larger for KL and KLP than MFI at θ s near 0. Differences decreased as test length increased. Item overlap was 0.00 for the 1 item test, but increased as the test-length increased, to .8 on the 20 item test. Chang and Yang (1996) found average bias and MSEs for item selection using KL than MFI at extreme negative θ values. This was most pronounced in the early stages; when there were more than 30 items, there were not differences between the methods.

In polytomous CATs, studies have found KL information methods perform comparably to MFI. (Lima Passos et al., 2007; Veldkamp, 2003). Lima Passos et al. (2007) found comparable performance in terms of RMSE and bias using the NRM on a 15 item test. Similarly, Veldkamp (2003) found MFI and KL information methods had high overlap and comparable MSEs using the GPCM and a 20 item test.

Gradual Maximum Information Ratio.

Han (2009) proposed the gradual maximum information ratio (GMIR) as an alternative to MFI criterion. GMIR uses the ratio of the expected information at the interim θ to the potential maximum information. Through a weight function, GMIR uses this ratio of item efficiency for item selection in the earlier stage of the CAT and then places more importance on the MFI, or item effectiveness, towards the end of the CAT. GMIR selects the item that maximizes the equation

$$\frac{\boldsymbol{I_i}[\hat{\theta}_{m-1}]}{\boldsymbol{I_i}[\theta^*]}\left(1-\frac{m}{M}\right)+\boldsymbol{I}\big[\hat{\theta}_{m-1}\big]\frac{m}{M} \qquad (17)$$

where $\hat{\theta}$ is the interim θ estimate, m-1 is the number of items administered so far, θ * is a the point of the θ scale where the maximum Fisher information is provided by item *i*, and M is the test length. $\frac{\boldsymbol{I_i}[\hat{\theta}_{m-1}]}{\boldsymbol{I_i}[\theta^*]}$ is the item efficiency ratio, or the ratio of the information at the interim estimate as compared to the potential peak of the information function. θ * is equal to b_i when the c-parameter is equal to zero. The ratio of $\frac{m}{M}$, the location within the test, determines if item efficiency or item effectiveness receives more weight in the item selection calculation. In the early stages of the CAT, when a few items have been administered, $\left(1-\frac{m}{M}\right)$ is larger than $\frac{m}{M}$ and the item efficiency ratio is given more weight. As the test progresses, $\frac{m}{M}$ becomes larger and the item effectiveness receives more weight.

While previous studies comparing GMIR and MFI, have not found meaningful differences in precision of measurement of final θ estimates on 10-40 item tests (Chang & Dodd, 2013; Han, 2009; Han, 2010), these studies have shown slightly better measurement precision when GMIR was used. Chang and Dodd (2013) found differences in standard error of estimate, bias and RMSE. These differences were greater in magnitude when the CATs were shorter, at the early stages of the CAT, and at extreme θ estimates. In the early stage of the CAT, or items 1-5 or 1-10, the GMIR places emphasis on the item efficiency and then places more importance on the item effectiveness on subsequent items. By the end of the CAT, GMIR and MFI will both select items with maximum effectiveness, or GMIR is essentially the MFI criterion at this point. When the items selected by the item selection methods overlap greatly, the performance should be similar (Chen et al., 2000; Veldkamp, 2003). However, over the first few items, when the item selection methods place emphasis on different aspects of the items, selection of different items between GMIR and MFI, performance should vary.

In a dichotomous CAT, Han (2009, 2010) found no meaningful differences in performance of θ estimation between MFI and GMIR item selection methods. Using a 3PL model in a constrained 40 item CAT, Han (2009) found the GMIR did result in a slightly lower standard error of θ estimates as compared to MFI. The mean standard errors of estimates were smaller across 20 days of administration. However, this difference was small and not meaningful. Standard error of estimates, mean absolute error, and bias were similar across the θ scale for MFI and GMIR. Expanding to 10 item, 20 item, and 40 item CATs, Han found MFI, GMIR, and KL item selection methods outperformed IIC, LWI, and a-stratified procedures. MFI, GMIR, and KL had lower standard error of estimates. Differences among the item selection procedure diminished for longer tests. Han (2009, 2010) also evaluated item pool usage and found more balanced item pool utilization using the GMIR.

In a polytomous CAT, Chang and Dodd (2013) found using the GPCM that GMIR and MFI had comparable overall measurement precision and administration efficiency. Both item selection methods had grand means and standard deviations near the values of the known θs, grand mean of 0.00 and standard deviation of 1.00. Additionally, the grand mean standard errors of the final θ estimates were identical. However, GMIR had a smaller mean standard error, mean bias, and mean RMSE than

MFI at the early stages of the CAT. The largest differences occurred at items two and three. GMIR also had smaller mean standard error, mean bias and mean RMSE at extreme known θ values, especially extreme negative θ values. Chang and Dodd also found that GMIR produced fewer nonconvergent cases than MFI. Across conditions, on average, MFI produced over five times as many nonconvergent cases as GMIR. They found the biggest differences in performance between MFI and GMIR occurred in the variable conditions at items 2 and 3. Varying the population distribution between normal and negatively skewed did not interact with the item selection method under the conditions of the study.

Bayesian Approaches

As stated previously, Bayesian item selection approaches chose items based on the prior and posterior distributions of θ estimates. Owen (1975) proposed the first Bayesian approach, an approximate Bayesian criterion based on a normal approximation of the posterior distribution. His criterion was a normal approximation because the fully Bayesian approach was too numerically complex for the time. Owen selected the first item based on the mean of a normal prior distribution. This prior is then multiplied by the likelihood of the observed response to find the posterior distribution. The next item selected is the item that minimized the expected variance of the posterior variance θ distribution, given the interim θ. Owen repeatedly approximated a new posterior and used it as the new prior for the next item. This process is stopped when the variance is smaller than a pre-specified value (van der Linden, 1998). Van der Linden (1998) introduced a number of

Bayesian approaches using the true posterior distribution as alternates to the approximate used by Owen (1975). These approaches are described below.

Maximum Posterior Weighted Information. The maximum posterior weighted information (MPWI) criterion was proposed as an alternative to maximum information (van der Linden, 1998). The MPWI weights the observed information function by the posterior θ distribution to account for the uncertainty of the interim θ estimate. MPWI uses different weights across θ levels, determined by the posterior θ distribution, as opposed to considering only one specific θ or having a uniform weight across a confidence interval. MPWI selects the item that maximizes the expected value of the observed information over the posterior distribution, or

$$\int J_{U_j}(\theta) g\left(\theta \middle| u_{i_1}, \ldots, u_{i_{k-1}}\right) d\theta; j \in R_k \tag{18}$$

where $J(\theta)$ is the observed information measure, $g\left(\theta \middle| u_{i_1}, \ldots, u_{i_{k-1}}\right)$is the posterior distribution, k is the location of the item on the test, and R_k is the set of unused items in the item pool.

Maximum Expected Information. Van der Linden (1998) also proposed the maximum expected information criterion (MEI), which weights the observed information measure by a posterior predictive distribution. Initially, the expected probability distribution for each response category U, conditional on θ, is calculated for each item $j \in R_k$. He then calculated the posterior predictive distribution, $p_j(U_j=u_j|u_{i_1}, \ldots, u_{i_{k-1}})$, using this expected probability distribution, $p_j(U_j=u_j| \theta)$, and the posterior distribution, $g\left(\theta \middle| u_{i_1}, \ldots, u_{i_{k-1}}\right)$. MEI selects the item that maximizes the expected information over the probability distribution of the examinee's responses on each item $j \in R_k$, or

$$pj\left(Uj = 0\middle|u_{i_1}, \dots, u_{i_{k-1}}\right)\mathrm{J}_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}\left(\hat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}\right) + \qquad (19)$$

$$pj\left(Uj = 1\middle|u_{i_1}, \dots, u_{i_{k-1}}\right)\mathrm{J}_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}\left(\hat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}\right); j \in R_k$$

The observed information measures $\mathrm{J}(\hat{\theta})$ differ according to the θ estimate based on the item response.

Minimum Expected Posterior Variance. The minimum expected posterior variance (MEPV) criterion, also introduced by van de Linden (1998), uses the posterior variance of θ for each category response, instead of the observed information measures, like the MEI. So, the criterion selects the item that minimizes the equation

$$pj\left(Uj = 0\middle|u_{i_1}, \dots, u_{i_{k-1}}\right)Var\left(\hat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}\right) + \qquad (20)$$

$$pj\left(Uj = 1\middle|u_{i_1}, \dots, u_{i_{k-1}}\right)Var\left(\hat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}\right); j \in R_k$$

Van der Linden proposed the MEPV criterion as a small-sample alternative to the MEI, since the reciprocal of the information is only an approximation to the true variance of the posterior in large samples.

Maximum Expected Posterior Weighted Information. The maximum expected posterior weighted information (MEPWI) criterion is a combination of the MPWI and MEI (van der Linden, 1998). The MEPWI uses the integral used in the MPWI instead of the observed information measure used in the MEI. This integral differs based on the item responses. So, the MEPWI criterion selects the item that maximizes the equation

$$pj\left(Uj = 0 \middle| u_{i_1}, \dots, u_{i_{k-1}}\right) \int \mathrm{J}_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}\,(\theta)\ \mathrm{g}\left(\theta \middle| u_{i_1}, \dots, u_{i_{k-1}}, U_j = 0\right) d\theta + \tag{21}$$

$$pj\left(Uj = 1 \middle| u_{i_1}, \dots, u_{i_{k-1}}\right) \int \mathrm{J}_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}\,(\theta)\ \mathrm{g}\left(\theta \middle| u_{i_1}, \dots, u_{i_{k-1}}, U_j = 1\right) d\theta;$$

$$j \in R_k$$

In a comparison of the four Bayesian methods using a 2PL model, van der Linden (1998) found the MEI, MEPV, and MEPWI had the best function, lower bias and MSEs, with MPWI next and MFI performing the worst, especially with fewer than 20 items and even after 30 items. He claimed that the use of the posterior predictive distribution of the item responses was the critical feature. Using the posterior only for weighing observed information was not enough and additional weighing of observed information did not result in benefits in efficiency.

Polytomous studies have shown varying results. Using the PCM, Penfield (2006) found MEI and MPWI performed comparably and outperformed MFI. MEI and MPWI had smaller RMSE and administered fewer items at extreme θ values than MFI, using a flat and peaked item bank. Differences were more pronounced using the peaked item bank. However, these differences were not found when θ = 0. In contrast, Choi and Swartz (2009) found Bayesian methods and MFI performed similarly, using the GRM and EAP θ estimation. They showed that no method performed better than other in terms of bias, RMSE, and correlations with true θ values. The small but not meaningful differences that did exist diminished as test length increased from 5 to 10 to 20 items. Similarly, item overlap was high across the methods, increasing as test length increased. Ho (2010) also found no practical difference among MFI and Bayesian approaches, using GPCM and multiple θ estimation methods: MLE, WLE, EAP with normal prior distribution, and EAP with positively skewed prior distribution. Ho found comparable mean bias, RMSE, absolute average difference, and intercorrelations of θ values.

Trait Estimation

Computer adaptive testing begins with an initial trait estimate for the test taker. The initial estimate can be the same for all test takers, the mean of the population distribution, or it can be based on prior information. An item is selected to maximize information at this initial θ estimate. Next, based on the response to the administered item, the test taker's trait estimate is updated. A second item is selected to maximize information at this new θ estimate. This process continues updating the θ estimate after every selected item until the stopping rule is reached and the test is terminated. The final trait estimate is ultimately calculated based on all the examinee's responses to the administered items.

CAT trait estimates are generally based on either the likelihood function or the posterior distribution. The likelihood function is the likelihood of the set of responses to the items for the range of θ values, or $L(\theta \mid x_n)$, where x_n is the set of item responses. It is calculated as the product of the conditional probabilities of the item responses. The posterior distribution is the likelihood function times the prior distribution, the hypothesized population distribution from which the examinees are sampled (Embretson & Reise, 2000).

CAT studies have used the following estimation procedures: maximum likelihood estimation (MLE) and Warm's (1989) weighted likelihood estimation (WLE), based on the likelihood function; and expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) and maximum a posteriori (MAP) estimation (Bock, 1983), based on the posterior distribution.

The maximum likelihood estimate is the θ value where the likelihood function has the highest value, or the maximum of *L* from the likelihood function. This θ value is located using a Newton-Raphson iteration. The WLE procedure weights the θ estimate and can reduce the standard error and bias in fixed-length CATs, but does not reduce the bias and standard error in variable-length CATs (Boyd et al., 2010, Wang, Hanson & Lau, 1999; Warm, 1989). Both the MLE and WLE can be estimated after the first item in a polytomous CAT, if the response is not in an extreme category. Similarly, MLE and WLE can be estimated in a dichotomous CAT once there are correct and incorrect item responses. Until a θ value can be estimated, studies use a fixed- or variable-step-size procedure.

The variable-step-size procedure is the most commonly used. The variable-step-size procedure generally works better than the fixed-step-size procedure, resulting in fewer nonconvergent cases (Dodd, 1990). For a dichotomous CAT, the interim θ estimate is calculated as half the distance from the initial θ to the smallest or largest item difficulty value, for an incorrect or correct item respectively. In polytomous CAT, the interim θ estimate is calculated as half the distance from the initial θ to the lowest or highest step difficulty value, for the lowest or highest category respectively. This procedure occurs until a dichotomous CAT has at least one correct and one incorrect response, or for a polytomous CAT, there are responses in two different categories (Dodd, 1990; Dodd et al., 1995).

Both EAP and MAP are Bayesian procedures based on the posterior distribution. The EAP estimate is the θ value associated with the mean of the posterior distribution, located through a noniterative procedure. The MAP estimate is the θ value associated with the mode of the posterior distribution, located through Newton-Raphson iteration. Bayesian procedures can be estimated after the first item, regardless of the response category.

Bayesian procedures tend to regress toward the mean of the prior distribution (Baker & Kim, 2004; Bock & Mislevy, 1982; Weiss, 1982). Also, incorrect or inappropriate prior distributions can bias the estimates, especially for short tests (Mislevy & Stocking, 1989; Seong, 1990). The impact of the prior distribution is greater in the early stages and decreases as more items are administered (Wainer, 2000). However, Bayesian procedures will always result in θ estimates, unlike MLE and WLE, in which the Newton-Raphson iterations may not always converge (Embretson & Reise, 2000). Bayesian procedures also generally decrease the standard errors associated with the trait estimates.

Based on a 3PL model, Wang and Vispoel (1998) found MLE produced lower bias, higher standard errors, higher RMSE, lower fidelity, and lower administration efficiency. Recently, in a comparison of MLE, EAP with a normal prior, and EAP with a uniform prior estimation in a polytomous CAT based on the RSM, Chen, Hou, Fitzpatrick, and Dodd (1997) found MLE and EAP performed comparably in terms of θ estimation. All three estimation methods performed comparably with a normal and a negatively skewed population distribution. Performance of EAP was comparable even when the prior did not match the underlying trait distribution. Additionally, using EAP with a normal prior, regression to the mean occurred at the extremes, but was minimal and not practically important.

Stopping Rule

Options for CAT termination are variable length method and fixed length method. Additionally, a combination of these options can be used. The two mostly commonly used stopping rules are the variable length and fixed length (Boyd et al., 2010; Dodd et al., 1995; Wainer, 2000).

Fixed-Length Test

The fixed length method terminates a CAT when a pre-specified number of items are administered, regardless of the standard error. Advantageously, it is easy to explain since all participants take the same number of items, like a conventional test. However, this method does not measure individuals at the same level of precision. Typically measurement errors are larger at the extreme θ s than at the middle θ s, where the typical item bank has more informative items. Measurement error for a given θ is proportional to the number of items with difficulty parameters matching that θ (Wainer, 2000). Another disadvantage is that for some examinees, certain items may not contribute much information about their trait level, which is not efficient testing.

Variable-length Test

Variable length CATs seek a specific measurement precision, stopping administration when either a pre-specified standard error or minimum information rule is met, regardless of the number of items administered (Dodd et al., 1993). The most commonly used, the standard error stopping rule, terminates a CAT once a pre-specified standard error value, either a universal or conditional on the trait estimate, is reached (Boyd et al., 2010; Wainer, 2000). The standard error stopping method is common is medical outcome research (Ware et al., 2000; Ware et al., 2005; Ware et al., 2003). In medical outcome research, a standard error value that is conditional on the trait level allows for different precision across the range of trait estimates. Individuals with more severe medical problems, higher θ estimates, can be measured with more precision than those with less severe medical problems. Conversely, a higher standard error stopping rule can be used for individuals with very low θ estimates who are not clinically relevant (Ware et al., 2003). While standard error methods measure individuals at the same predetermined levels of precision, the number of items administered to individuals will vary. It may not appear to test takers that they were measured accurately and/ or with enough items.

Administering a minimum number of items can prevent this issue (Gershon, 2005). The standard error stopping rule may administer additional items even when the predetermined standard error cannot be met, which can place undue burden on the examinee. This could be especially undesirable for a brief medical screening (Gardner et al., 2004). Conversely, the standard error stopping rule might limit the measurement precision to the predetermined SE by stopping the CAT when additional informative items are available.

Using the minimum information stopping method, a CAT is terminated when the items remaining in the item pool offer less than the minimum level of information, at the current θ estimate. The minimum information method prevents unnecessary administration of items that contribute little information. However, studies have found it resulted in too few items administered and therefore a large number of nonconvergent cases using the NRM and GRM (De Ayala, 1989; Dodd, Koch, & de Ayala, 1989). Also, studies have found that among the variable length stopping rules, the standard error method performs better than the minimum item information method with respect to the measurement precision, mean number of items administered, correlation of known and estimated traits, and number of nonconvergent cases using the NRM (De Ayala, 1992), the RSM (Dodd, 1990), the PCM (Dodd et al., 1993), and the GRM (Dodd et al., 1989). Using the Rating Scale Model, Dodd (1990) found the minimum information stopping rule resulted the CAT terminating after three to four items for extreme θ, when an item with the minimum information was not found, and ultimately in high standard errors.

The variable length stopping rule can be used in conjunction with the fixed length stopping rule; a test will terminate after a pre-determined number of items if the standard error was not met. This combination is often applied in practice (Wainer, 2000). This combination of methods can prevent an individual from the burden of a large number of items and the test from running out of items before the pre-determined precision level is reached. However, this combination of variable and fixed length can limit the precision of measurement, or the SE from decreasing below the predetermined value.

Research of PRO measures generally study tests with 5-10 items. Ware, (2005) used 5 item and 10 item rehabilitation outcome measures in their simulation study comparing these two CAT measures to a static survey. They suggest future studies expand beyond the 5 and 10 item conditions, "to better understand the optimal length of a dynamic CAT. Cook et al. (2007), in a simulation study of a 15-item general distress pool, used a fixed stopping rule of 7 items and a variable stopping rule with a 0.5 standard error cutoff, resulting in 8 items on average using PCM and 11 using GPCM. Fliege et al. (2005) used a .32 standard error cutoff, which resulted in 6 items on average for -2 θ to +2 θ, in a simulation study comparing the Depression-CAT to the full Beck Depression Inventory and an 8 item CES-D short form. Ware et al. (2003) compared a 6 item CAT to a 6 item short fixed form and a 54 item total test of a Headache Impact Test in a simulation study.

STATEMENT OF PROBLEM

While MFI is still the most commonly used item selection method for CAT due to its effectiveness and ease of use, alternative methods continue to try to overcome the attenuation paradox drawback of MFI. The attenuation paradox, which is especially problematic in the early stages of CAT, is particularly concerning in health care patient reported outcome measures where fewer items can relieve patient burden. Overall, most proposed alternatives to MFI had comparable or poorer performance to MFI in most studies using polytomous CATs, especially on shorter tests (Choi & Swartz, 2009; Ho, 2010; Lima Passos et al., 2007; van Rijn et al., 2002; Veldkamp, 2003). Han (2009,2010) surmised GMIR selecting the most efficient item could be more robust than the MFI selecting the most effective item against the interim Θ instability in the early stages of CAT. While Han did not find meaningful differences between MFI and GMIR, using dichotomous models, he did find GMIR resulted in a smaller standard error of estimate over a 40 item CAT.

Similar to previously developed item selection methods, GMIR was developed and initially studied with dichotomous items, under the 3PL model. Little research has considered this item selection method with polytomous items. The previous polytomous research showed that GMIR may perform better than MFI at extreme trait values and in the early stages of the CAT. Both of these situations could be especially advantageous in patient reported outcome measures. No existing research has investigated GMIR under conditions similar to PRO measures, with very few items, only 5-10 items. Also, previous research has only compared MFI and GMIR selection methods using constrained CATs, employing content balancing and/or exposure control. These conditions are not realistic and applicable for most medical outcome measures. The previous polytomous research used the generalized partial credit model, but for an attitude measure like a PRO, Andrich's Rating Scale Model (1978) is more appropriate. This model was developed for Likert-type attitude scales and is more parsimonious.

This study will investigate the following research questions:

1) How do the MFI and GMIR item selection methods' performances compare for CATs with small numbers of items?

2) How do the item pool size, mismatch of item pool and population latent trait distributions, and test length affect the measurement precision?
3) How do interactions among item pool size, mismatch of item pool and population latent trait distributions, and test length affect the MFI and GMIR item selection methods' performance?

Chapter 3: Methodology

DESIGN OVERVIEW

This CAT simulation study compares the performance of the MFI and GMIR item selection methods under ARSM. Performance was evaluated in terms of measurement precision and item administration efficiency. Four independent variables were manipulated: item selection method, item pool size population distribution, and test length. The resulting design is a 2 x 2 x 2 x 3 factorial design.

Only MFI and GMIR item selection methods were compared because MFI has greater ease of use, and previous research with polytomous CATs has shown other item selection methods do not perform better than MFI (Choi & Swartz, 2009; Ho, 2010; Lima Passos et al., 2007; van Rijn et al., 2002; Veldkamp, 2003). The MFI and GMIR performance was compared under two population latent trait level distributions, one that is a match to the item pool distribution—a normal distribution—and one that is a mismatch—a negatively skewed distribution. This mismatch was included because previous research with polytomous CATs has found less accurate trait estimates when there is a mismatch between the examinees' trait levels and item pool distributions (Dodd et al., 1993; Gorin et al., 2005). In health measures, item pool and examinee trait levels might not match if the subpopulation levels do not align with the general population or levels might match pretreatment, but not match post treatment. Item selection methods were also evaluated under two different item pool sizes. One item pool consisted of 41 polytomous items and the other consisted of 82 polytomous items. Previous studies have shown that polytomously scored item pools can have as few as 30 items (Dodd 1990; Dodd & de

Ayala, 1994). The stopping rule was a combination of variable-length and fixed-length. The CAT stopped either when the standard error of measurement of the simulee's θ estimate was less than the prespecified standard error (0.40, 0.45, or 0.54) or when a set number of items had been administered (5, 7, or 9).

With a 2 x 2 x 2 x 3 factorial design, the study has 24 conditions total. Each condition has 1000 simulees and was replicated 100 times. Variable step size and MLE were used to estimate trait levels.

ITEM POOL AND TEST CHARACTERISTICS

An existing database was used to perform real-data simulations. The item pool that was used in this study consists of 41 polytomous items, the developmental form of the flexilevel shoulder functioning scale (Cook, Roddey, Gartsman, & Olson; 2003). Each item has five response categories, with higher score signifying better shoulder function. The response categories and corresponding scores are "no difficulty" 4, "little difficulty" 3, "some difficulty" 2, "much difficulty" 1, "I can't do this" 0, and "didn't do before shoulder problem" not applicable. Dodd, Cook, and Godin (2010) calibrated the 41 items according to the partial credit model (Master, 1982) using PARSCALE 4. Dodd et al. (2010) stated that they obtained the estimates of the threshold and scale values for Andrich's rating scale model (ARSM) from the PARSCALE output using the procedures of Wright and Masters (1982). Item responses were collected from 400 participants, who were recruited at 3 facilities: an orthopedic surgeon's office, a county Physical Therapy department, and the Houston Veteran's Affairs Medical Center Hospital. Using a procedure previous used in CAT simulation studies (Dodd, Koch, & de Ayala, 1993), the item parameters for the 82 items were calibrated. The item pool was expanded to 82 by doubling the original 41 item parameters. A calibration sample was generated using simulated data for the doubled item pool, to introduce variability. These 82 items were calibrated with PARSCALE 4. Responses from 1000 normally distributed simulees were generated according the ARSM using the IRTGEN SAS macro program (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). Using these responses, new threshold and scale value estimates were obtained from PARSCALE using the procedures of Wright and Masters (1982). The step values were estimated using standard marginal maximum likelihood estimation procedures. Specifically, for the scale value, the step parameter estimates were averaged for the partial credit model; for the thresholds, the deviation of the step parameter estimates from the scale value across items were averaged. Descriptive statistics for item parameter estimates are presented in *Table 1*. Scale values ranged from -2.14 to 1.32, and their average was slightly negative (-0.64). An information plot for the item pool calibrated according to the ARSM is presented in *Figure 3*. The first 41 items were used for the 41 item item pool conditions.

	Mean	SD	Minimum	Maximum
Scale Value	-0.64	.73	-2.14	1.32
Thresholds				
1	-1.40			
2	-0.75			
3	0.68			
4	1.46			

Table 1. Descriptive Statistics for the Item Parameter Estimates for the Shoulder Functioning Data

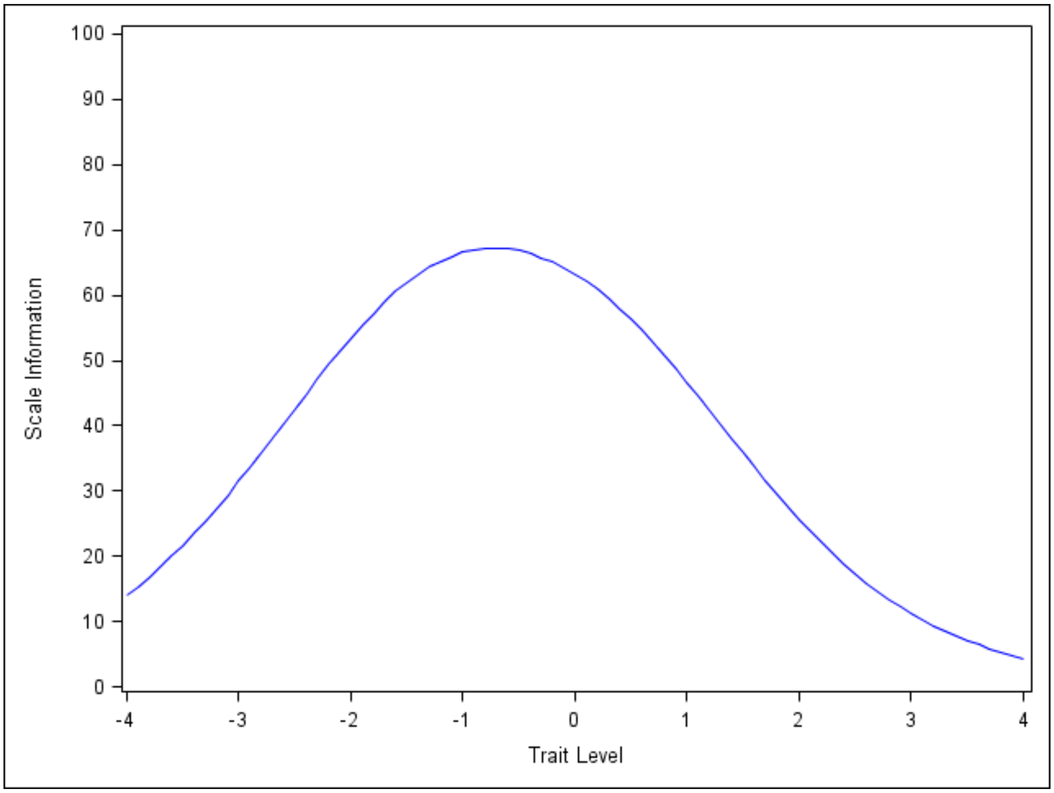


Figure 3. Information Function for the Shoulder Functioning Scale 82-Item Item Pool Calibrated According to Andrich's Rating Scale Model

DATA GENERATION

For the condition in which the latent trait distribution of the population matches the item pool distribution, 1000 θ values were randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 1. This was replicated 100 times to create 100 samples. For the condition in which the trait distribution of the population is a mismatch to the item pool distribution, a negatively skewed population distribution was simulated. The normal and skewed population distributions were used since these are encountered most frequently in practice. Since a positively skewed distribution is the mirror image of the negatively skewed distribution, only one is used. Using the procedure used in previous polytomous CAT simulation studies (Davis, 2004; Davis & Dodd, 2003; Gorin et al., 2005; Koch & Dodd 1989), the parameters of a beta distribution were set, α to 5.0 and β to 1.8, in order to obtain the negatively skewed shape of the distribution. From this negatively skewed distribution, 1000 θ values for each of the 100 replications were randomly drawn. One of the negatively skewed distributions is presented in Figure 4. The skew for the sample distribution shown in Figure 4 was -0.66. The normally distributed and negatively skewed distributions were then standardized to have a mean of 0 and standard deviation of 1.

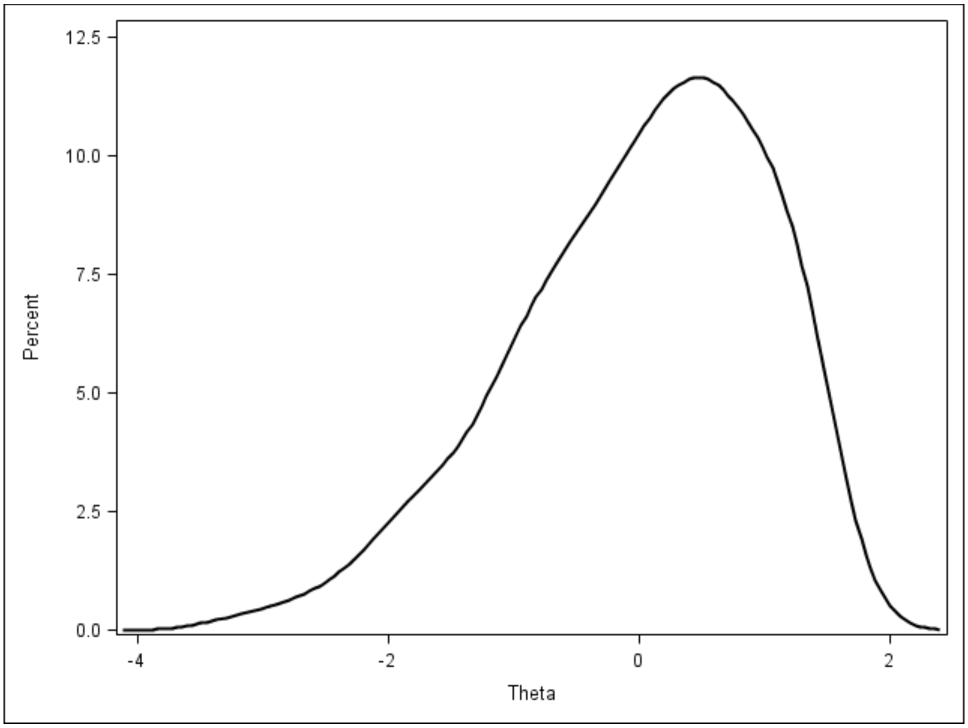


Figure 4. Negatively Skewed Theta Distribution for One Sample

Item responses based on Andrich's Rating Scale Model were generated using the IRTGEN SAS macro program (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). For each simulee, given the known θ and the item parameter estimates, the probability of responding in each category score was calculated (see Equation 4). These probabilities were summed to obtain cumulative subtotals for each category score, category score boundaries. These subtotals or boundaries are then compared to a number randomly drawn from a uniform distribution from 0 to 1. If the randomly drawn number was less than or equal to the category boundary, the simulee was assigned that category score. This comparison was successive from the low category boundary to the high category boundary. These comparisons were repeated for all items, all simulees, and all samples until all simulees had scores on all items. These 100 generated-response data sets were used for each CAT simulation.

CAT SIMULATION

For each condition, item response data and item pool characteristics were input into a CAT program, which is an adaptation of the program created by Dodd, Cook, and Godin (2005) for their study assessing the RSM and SIM with a medical assessment measure. Items were selected using two item selection procedures, MFI and GMIR. While MFI is the default in the program, for the GMIR conditions, the MFI item selection procedure was altered to include the ratio of item efficiency.

The item pool consisted of either the 41 or 82 items of the flexilevel shoulder functioning scale. The initial θ estimate was set to 0 for all simulees in all conditions and used to select the first item for administration. A variable step size approach, which sets the interim θ estimate at half the distance between the previous θ estimate and the extreme value in the item pool is recommended for polytomous CAT (Koch & Dodd, 1989; Dodd, 1990). Once there are responses in two different categories, maximum likelihood estimation procedure was used to update the estimate of shoulder functioning for all subsequent item responses.

The CAT simulations was terminated with a combination of variable-length and fixed-length stopping rules. The CAT stopped either when the standard error of measurement of the simulee's θ estimate was less than the prespecified standard error or when a maximum number of items had been administered. Research of patient reported outcome measures generally study tests with 5-10 items (Cook et al., 2008; Cook et al., 2007; Fliege et al., 2005; Ware et al., 2005). The CAT stopped at a maximum of 5, 7, or 9 items administered or when the standard error reached 0.40, 0.46, or 0.54 respectively. The standard error appropriate for each number of item administered was determined by finding the average item information per item (for the item pool from θ of -2.5 to 1.5). Using this average item information per item and each number of items to be administered, the standard errors were calculated using Equation 8. This combination stopping rule has performed well and is recommended for CATs using various polytomous IRT models (Dodd et al., 1995). Since Han's (2009) GMIR method was developed for use with a fixed-length CAT, the weighting ratio $\frac{m}{M}$ from equation (16) was replaced with $\frac{target\ SE}{interim\ SE}$. This modification was used by McClarty et al. (2006) to modify the progressive-restrictive exposure control procedure for use with variable length tests. The modified GMIR equation is

$$\frac{\boldsymbol{I_i}[\hat{\theta}_{m-1}]}{\boldsymbol{I_i}[\theta^*]}\left(1-\frac{target\ SE}{interim\ SE}\right)+\boldsymbol{I}[\hat{\theta}_{m-1}]\frac{target\ SE}{interim\ SE} \tag{22}$$

DATA ANALYSIS

The test performance of the MFI and GMIR was examined in terms of the number of nonconvergent cases, descriptive statistics, correlations, mean bias and root mean squared error (RMSE) statistics, and conditional standard error and mean bias statistics under each experimental condition. Also, the number of items administered was used to evaluate the item administration efficiency.

Within each condition, before calculating outcome measure, listwise deletion of nonconvergent cases was performed. A case was considered nonconvergent if MLE is never reached or the final ability estimate was too extreme (equal to either -4 or +4). The mean number of nonconvergent cases for each condition is reported. Cases of nonconvergence that never reach MLE estimation are reported as "nonconvergent". Cases, in which the final ability estimate is too extreme, are reported as "out of range". These two types of nonconvergent cases are reported separately for each condition. Each of the measures used to evaluate the conditions were averaged across the 100 replications. In application, fewer nonconvergent cases would mean that more patients receive PRO measure scores and assessments of their functioning levels. The two types of nonconvergent cases are separated because they differ in practical usefulness. A patient who received an out of range nonconvergent score would know that his functioning was extremely poor or extremely high, which provides information. However, a nonconvergent case that did not reach MLE would not provide any information of functioning level. It would be burdensome for a patient to spend the time to take a measure and not receive this assessment of his functioning level.

Descriptive statistics and comparison of the final θ estimate to known θ values were used to evaluate measurement precision. These statistics evaluate the recovery of known θ values. The grand mean, mean maximum, and mean minimum of the final θ estimates ($\hat{\theta}$), and their standard errors (SE) are reported for each condition. Mean, minimum, and maximum Pearson product-moment correlation between known and estimated θ values, mean bias, and root mean squared error (RMSE) statistics were calculated for trait estimates produced by each CAT condition. Bias is the average difference between the known and estimated θ levels and is an index of systematic error of measurement. RMSE is an index of the total error of measurement. The equation for bias is

$$\frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)}{n} \tag{23}$$

and the equation for RMSE is

$$\left[\frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)^2}{n}\right]^{\frac{1}{2}} \tag{24}$$

where $\hat{\theta}_k$ is the final trait estimate for patient *k*, θ_k is the known trait level of patient *k*, and *n* is the total number of patients.

These comparisons of the Pearson product-moment correlation between known and estimated θ values, mean bias, and RMSE across conditions using MFI and GMIR shows how the item selection methods perform in terms of measurement precision. Comparisons were made between different levels of the independent variables to determine how these affected the measurement precision of the item selection methods. To illustrate how the item pool size affects the measurement precision, conditions with 41 items and 82 items in the item pool were compared on these outcome variables. Similarly, a contrast of these outcome variables was made between conditions when the item pool and population latent trait distributions match as compared to a mismatch to show how distribution mismatching affects the measurement precision. Comparing outcome variables among the three test length conditions demonstrates how the test length affects the measurement precision. To determine how interactions among the item pool size, mismatch of item pool and population latent distribution, and test length affect the measurement precision, cell means of correlations, mean bias, and RMSE for each condition were compared.

Further, conditional plots of mean bias, mean RMSE, and grand mean standard errors were generated to illustrate the precision of the final and interim θ estimates across the range of θ values and the length of the test, respectively. For the final θ estimates, the known θ values of simulees were grouped into 0.5 intervals from -4 to +4 so that midpoints of the groups are spaced equally from -4 to +4. Outcome measures were calculated and plotted against the mean θ for each group, assessing the measurement precision across the range of trait values.

For the interim θ estimates, simulees were pooled into five groups with midpoints at θ = –2, –1, 0, 1, and 2, each group including those 0.5 standard deviations above and below. Outcome measures were calculated and plotted against item number, assessing the measurement precision at each item in the test. Additionally, the grand mean, mean minimum, and mean maximum number of items administered (NIA) are reported to evaluate the efficiency of the CAT, with smaller mean NIA signifying greater efficiency.

Chapter 4: Results

NONCONVERGENT CASES

After all conditions were run and before statistics were calculated, listwise deletion was performed. Two types of nonconvergent cases were deleted. Cases were considered out of range nonconvergent if the θ estimate was less than or equal to -4 or greater than or equal to 4. Cases were considered nonconvergent if MLE was not reached. Table 2 shows the mean number of out of range and nonconvergent cases for each condition averaged across the 100 replications. On average across replication and conditions, MFI resulted in fewer overall nonconvergent cases, especially cases that were out of the θ range from -4 to 4. GMIR and MFI simulations averaged 7.27 and 1.42 out of range examinees respectively and 2.35 and 1.95 nonconvergent cases respectively.

For out of range cases, GMIR simulations resulted in a larger range, averaging from 1.81 (5 item stopping rule/normal/41 item pool) to 11.18 (7 item stopping rule /negatively skewed/41 item pool). MFI simulations averaged from 1.09 (7 item stopping rule /normal/41 item pool) to 1.75 (9 item stopping rule /normal/41 item pool) out of range conconvergent cases. For nonconvergent cases that did not reach MLE, GMIR simulations averaged 1.00 (9 item stopping rule /negatively skewed/41 item pool) to 6.48 (5 item stopping rule / negatively skewed/82 item pool). MFI simulations averaged 1.00 (9 item stopping rule /negatively skewed/41 item pool) to 4.37 (5 item stopping rule / normal/ 41 item pool).

Both item selection methods had a greater number of nonconvergent cases that did not reach MLE when fewer items were administered. Conditions using a 5 item stopping rule averaged 4.593 and 3.43 cases not reaching MLE using GMIR and MFI respectively, whereas conditions using a 7 or 9 item stopping rule averaged fewer cases: 1.23 with GMIR and 1.21 with MFI. GMIR had a great number of nonconvergent cases not reaching MLE in the 5 item condition when the 82 item pool was also used as compared to the 41 item pool, 6.47 items as compared to 2.72. MFI resulted in the opposite trend, with greater nonconvergent cases when the 41 item pool was used with the 5 item stopping rule, 4.00 as compared to 2.87 with the 82 item pool. MFI resulted in fewer nonconvergent cases when the population was negatively skewed in the 5 and 7 item stopping rule conditions, with 3.72 and 1.32 cases as compared to 3.15 and 1.19 cases when the population was normally distributed. For practical purposes, this is a half a cases or less on average. GMIR conditions followed the same trend, but the differences were even smaller, around one tenth of a case on average.

In contrast, for nonconvergent cases that were out of the θ range, GMIR had fewer cases in the 5 item stopping rule condition and MFI did not vary across test length. GMIR simulations averaged only 2.603 out of range cases under 5 items, but 9.608 under 7 and 9 items. GMIR had more out of range cases in the conditions with the negatively skewed population distribution that was a mismatch to the item pool distribution, averaging 2.95, 10.17, and 10.05 in the 5, 7, and 9 item conditions. In the normally distributed population conditions, there were only 2.26, 8.86, and 9.36 cases. MFI showed the same trend in the 5 and 7 item stopping rule conditions, with

1.30, and 1.38 out of range cases in the normal condition and 1.52 and 1.47 in the negatively skewed condition. In the 9 item stopping rule condition, the trend was reversed, with 1.59 cases in the normal condition and only 1.27 cases in the negatively skewed condition. However, all these differences in the MFI conditions are too small to be practically important.

Condition			Out of Range			Nonconvergent		
			Mean	Min	Max	Mean	Min	Max
Normally Distributed Population	41 Item Pool	5 item						
		GMIR	1.81	1	5	2.88	1	7
		MFI	1.32	1	3	4.37	1	11
		7 item						
		GMIR	9.80	3	16	1.35	1	3
		MFI	1.09	1	2	1.43	1	3
		9 item						
		GMIR	9.72	4	15	1.19	1	2
		MFI	1.75	1	4	1.14	1	2
	82 Item Pool	5 item						
		GMIR	2.70	1	7	6.45	1	14
		MFI	1.28	1	4	3.06	1	7
		7 item						
		GMIR	7.91	1	14	1.48	1	5
		MFI	1.67	1	4	1.20	1	3
		9 item						
		GMIR	9.00	2	15	1.19	1	2
		MFI	1.43	1	4	1.06	1	2
Negatively Skewed Population Distribution	41 Item Pool	5 item						
		GMIR	2.73	1	7	2.56	1	7
		MFI	1.54	1	5	3.62	1	10
		7 item						
		GMIR	11.18	2	21	1.26	1	3
		MFI	1.51	1	4	1.22	1	4
		9 item						
		GMIR	9.82	2	20	1.00	1	1
		MFI	1.25	1	3	1.00	1	1
	82 Item Pool	5 item						
		GMIR	3.17	1	9	6.48	3	13
		MFI	1.49	1	3	2.67	1	7
		7 item						
		GMIR	9.16	3	18	1.30	1	3
		MFI	1.42	1	4	1.16	1	2
		9 item						
		GMIR	10.27	4	21	1.07	1	2
		MFI	1.29	1	4	1.50	1	2

Table 2. Nonconvergent Cases Averaged Across 100 Replications

ESTIMATED THETAS

Descriptive statistics illustrate how well the known θ values are reproduced. The grand mean and the mean of standard deviations of estimated θs averaged across 100 samples are in Table 3. The average minimum and maximum of the estimated θs as well as the mean, minimum, and maximum of the standard errors are also in Table 3. The mean estimated θs ranged from -0.031 to -0.015 with GMIR and -0.082 to -0.014 with MFI, with the minimum means in the 5 item stopping rule/ normal/82 item pool condition and the maximum means in the 9 item stopping rule / negatively skewed/ 41 item pool. These are all slightly lower than the known θ mean of 0.0. Conditions using the 5 item stopping rule and MFI had slightly lower average estimated θs (-0.074) than conditions using 7 and 9 item stopping rules and MFI (-0.022). GMIR was more consistent across stopping rules, with θ estimates all around -0.02. The θ estimates did not vary by item pool size, population distribution, or interactions among these conditions.

The mean standard deviations of estimated θs ranged from 1.084 to 1.152 with GMIR in the 9 item stopping rule /negatively skewed/41 item pool and 5 item stopping rule /normal/82 item pool respectively. With MFI, the standard deviations ranged from 1.079 to 1.161 in the 9 item stopping rule /negatively skewed/82 item pool and 5 item stopping rule /normal/41 item pool respectively. These are all slightly higher than the standard deviation for the known θs of 1. For both item selection methods and across distributions and item pool conditions, the standard deviation decreased slightly as the items administered increased, getting closer to 1. The overall average standard deviations for the 5,7, and 9 item conditions were

1.147, 1.104, and 1.084 respectively. Standard deviations did not vary by item pool size, population distribution, or any interaction among these variables.

Illustrating the measurement precision, the standard errors (SE) of the estimated θs ranged from 0.397 to 0.532 with GMIR in the 9 item stopping rule / negatively skewed /82 item pool and 5 item stopping rule /normal/82 item pool conditions respectively. MFI mean SEs ranged from 0.389 to 0.514 in the 9 item stopping rule / negatively skewed /82 item pool and 5 item stopping rule /normal/41 item pool conditions respectively. Across all conditions with a stopping rule of an SE of 0.54 or 5 items, the grand mean SE was 0.529 with GMIR and 0.511 with MFI. Across all conditions with a stopping rule of 0.46 or 7 items, the grand mean SE was 0.457 with GMIR and 0.449 with MFI. Across all conditions with a stopping rule of an SE of 0.40 or 9 items, the grand mean SE was 0.399 with GMIR and 0.391 with MFI. There were not differences between the two population distributions or the two item pool size simulations. There were also not interactions among these variables.

Condition			Final θ Estimate			Standard Error		
			Grand Mean(SD)	Min	Max	Mean	Min	Max
Normal Pop. Dist.	41 Item Pool	5 item						
		GMIR	-0.025(1.140)	-3.343	3.735	0.528	0.482	1.022
		MFI	-0.072(1.161)	-3.743	3.708	0.514	0.481	1.023
		7 item						
		GMIR	-0.019(1.110)	-3.702	3.912	0.458	0.414	1.015
		MFI	-0.021(1.109)	-3.704	3.933	0.452	0.408	1.015
		9 item						
		GMIR	-0.022(1.085)	-3.676	3.992	0.401	0.368	1.013
		MFI	-0.019(1.084)	-3.988	3.353	0.394	0.360	0.861
	82 Item Pool	5 item						
		GMIR	-0.031(1.152)	-3.760	3.867	0.532	0.481	1.066
		MFI	-0.082(1.155)	-3.928	3.949	0.511	0.482	1.020
		7 item						
		GMIR	-0.028(1.110)	-3.910	3.963	0.457	0.408	1.014
		MFI	-0.031(1.101)	-3.884	3.459	0.449	0.409	0.866
		9 item						
		GMIR	-0.025(1.090)	-3.967	3.640	0.397	0.360	0.894
		MFI	-0.023(1.083)	-3.827	3.648	0.391	0.360	0.866
Neg. Skewed Pop. Dist.	41 Item Pool	5 item						
		GMIR	-0.017(1.127)	-3.343	3.736	0.525	0.482	1.022
		MFI	-0.064(1.150)	-3.743	3.709	0.511	0.481	1.023
		7 item						
		GMIR	-0.019(1.102)	-3.702	3.912	0.456	0.414	1.015
		MFI	-0.018(1.097)	-3.704	3.933	0.449	0.409	1.015
		9 item						
		GMIR	-0.015(1.084)	-3.676	3.292	0.399	0.368	0.964
		MFI	-0.014(1.082)	-3.988	3.353	0.391	0.360	0.768
	82 Item Pool	5 item						
		GMIR	-0.026(1.145)	-3.760	3.867	0.530	0.481	1.066
		MFI	-0.079(1.143)	-3.928	3.949	0.508	0.482	1.020
		7 item						
		GMIR	-0.023(1.103)	-3.910	3.963	0.456	0.408	1.014
		MFI	-0.027(1.098)	-3.884	3.459	0.447	0.409	0.866
		9 item						
		GMIR	-0.022(1.086)	-3.967	3.207	0.397	0.360	0.894
		MFI	-0.020(1.079)	-3.827	3.648	0.389	0.360	0.866

Table 3. Estimated Thetas and Standard Error Descriptive Statistics Averaged Across 100 Replications

OVERALL MEASUREMENT PRECISION

The Pearson product-moment correlations are shown in Table 4. The correlations between the known and estimated θs show how well the known θs were recovered. Correlations range from 0.873 to 0.923 using GMIR in the 5 item stopping rule /negatively skewed/82 item pool and 9 item stopping rule/negatively skewed/82 item pool conditions respectively. Using MFI, correlations range from 0.856 to 0.922 in the 5 item stopping rule /negatively skewed/82 item pool and 9 item stopping rule/ negatively skewed/41 item pool conditions respectively. Conditions with more items administered had higher correlations. With the 5 item stopping rule, GMIR and MFI simulations resulted in correlations on average of 0.875 and 0.859 respectively. When the test length increased with a stopping rule of 7 items, correlations increased to 0.902 for GMIR and 0.899 for MFI. Correlations increased again to 0.923 with GMIR and 0.921 with MFI, when a stopping rule of 9 items was used. The correlation coefficients did not vary according to item pool size, population distribution or interactions among these variables.

The Bias and RMSE are shown in Table 5. These comparisons of the final θ estimates to the known θ values evaluate the overall measurement precision of the simulation conditions. Bias for simulations using GMIR ranged from 0.013 in the 7 item stopping rule /normal/41 item pool and 9 item stopping rule/ negatively skewed/ 41 item pool conditions to 0.033 in the 5 item stopping rule / negatively skewed/82 item pool condition. MFI simulation conditions resulted in a larger range of bias, from 0.015 in the 9 item stopping rule/ normal/ 41 item pool condition to 0.084 in the 5 item stopping rule /negatively skewed/ 82 item pool condition. Bias for MFI 5 item conditions was on average 0.077 as compared to 0.026 and 0.019 for 7 and 9 item conditions respectively. Bias for GMIR was similar across 5, 7, and 9 item stopping rule conditions and all smaller than MFI at 0.029, 0.019, and 0.016 respectively. For both MFI and GMIR, bias was slightly larger when the 82 item pool was used, especially when fewer items were administered. When 5 item stopping rule and 82 item pool were used, bias was on average 0.058 as compared to 0.048 with the 41 item pool. Using the 7 item stopping rule and 82 item pool, bias was on average 0.026 as compared to 0.018 with the 41 item pool. When the 9 item stopping rule and 82 item pool were used, bias was on average 0.02 as compared to 0.014 with the 41 item pool. However, these differences are too small for practical importance. Bias did not vary according to the population distribution, nor did the distribution interact with any other variables.

RMSE ranged from 0.418 in the 9 item stopping rule/ negatively skewed/ 41 item pool condition to 0.560 in the 5 item stopping rule/ negatively skewed/82 item pool condition with GMIR and 0.419 9 item stopping rule/ negatively skewed/ 82 item pool condition to 0.598 in the 5 item stopping rule/ negatively skewed/82 item pool condition with MFI. RMSE was larger for conditions with fewer items administered for both MFI and GMIR condition. RMSE values were slightly larger using MFI than GMIR, with greater differences with fewer item administered. MFI simulations resulted in RMSE values of 0.595, 0.483, and 0.424 for conditions with 5, 7,and 9 item stopping rules respectively. RMSE values of 0.554, 0.479,0.423 were found for GMIR conditions with stopping rules of 5, 7,and 9 items respectively.

RMSE values did not vary by item pool size or population distribution or interactions among these variables.

Condition			Correlation		
			Mean	Min	Max
Normally Distributed Population	41 Item Pool	5 item			
		GMIR	0.877	0.857	0.891
		MFI	0.862	0.847	0.880
		7 item			
		GMIR	0.903	0.887	0.914
		MFI	0.899	0.890	0.908
		9 item			
		GMIR	0.920	0.909	0.931
		MFI	0.921	0.906	0.930
	82 Item Pool	5 item			
		GMIR	0.875	0.855	0.892
		MFI	0.859	0.844	0.876
		7 item			
		GMIR	0.901	0.883	0.914
		MFI	0.898	0.885	0.914
		9 item			
		GMIR	0.922	0.912	0.932
		MFI	0.919	0.907	0.929
Negatively Skewed Population Distribution	41 Item Pool	5 item			
		GMIR	0.874	0.856	0.890
		MFI	0.859	0.839	0.872
		7 item			
		GMIR	0.901	0.889	0.916
		MFI	0.900	0.888	0.916
		9 item			
		GMIR	0.921	0.912	0.930
		MFI	0.922	0.910	0.931
	82 Item Pool	5 item			
		GMIR	0.873	0.854	0.887
		MFI	0.856	0.833	0.881
		7 item			
		GMIR	0.901	0.885	0.915
		MFI	0.899	0.885	0.917
		9 item			
		GMIR	0.923	0.912	0.934
		MFI	0.920	0.907	0.932

Table 4. Correlation Coefficient between Known and Estimated Thetas Averaged across 100 Replications

Condition			Bias			RMSE		
			Mean	Min	Max	Mean	Min	Max
Normally Distributed Population	41 Item Pool	5 item						
		GMIR	0.026	-2.814	3.037	0.549	0.508	0.591
		MFI	0.072	-2.744	3.183	0.593	0.557	0.630
		7 item						
		GMIR	0.013	-2.563	3.092	0.477	0.444	0.518
		MFI	0.021	-2.826	2.766	0.486	0.460	0.520
		9 item						
		GMIR	0.014	-2.644	2.758	0.426	0.399	0.454
		MFI	0.015	-2.245	2.820	0.424	0.397	0.450
	82 Item Pool	5 item						
		GMIR	0.030	-2.879	2.992	0.559	0.510	0.599
		MFI	0.083	-2.629	3.085	0.596	0.558	0.630
		7 item						
		GMIR	0.023	-2.463	3.215	0.482	0.447	0.519
		MFI	0.029	-2.109	3.074	0.486	0.458	0.519
		9 item						
		GMIR	0.018	-2.291	2.908	0.422	0.391	0.450
		MFI	0.022	-2.299	2.744	0.428	0.405	0.468
Negatively Skewed Population Distribution	41 Item Pool	5 item						
		GMIR	0.025	-2.879	3.029	0.549	0.517	0.587
		MFI	0.070	-2.931	3.349	0.593	0.555	0.631
		7 item						
		GMIR	0.016	-2.510	3.065	0.477	0.444	0.509
		MFI	0.021	-2.505	2.840	0.479	0.440	0.507
		9 item						
		GMIR	0.013	-2.138	2.732	0.424	0.394	0.448
		MFI	0.015	-2.228	2.736	0.419	0.392	0.458
	82 Item Pool	5 item						
		GMIR	0.033	-2.669	3.114	0.560	0.525	0.587
		MFI	0.084	-2.735	3.036	0.598	0.540	0.640
		7 item						
		GMIR	0.023	-2.571	2.969	0.480	0.440	0.517
		MFI	0.030	-1.856	2.931	0.482	0.437	0.517
		9 item						
		GMIR	0.018	-1.883	2.966	0.418	0.389	0.449
		MFI	0.022	-2.009	2.897	0.423	0.390	0.451

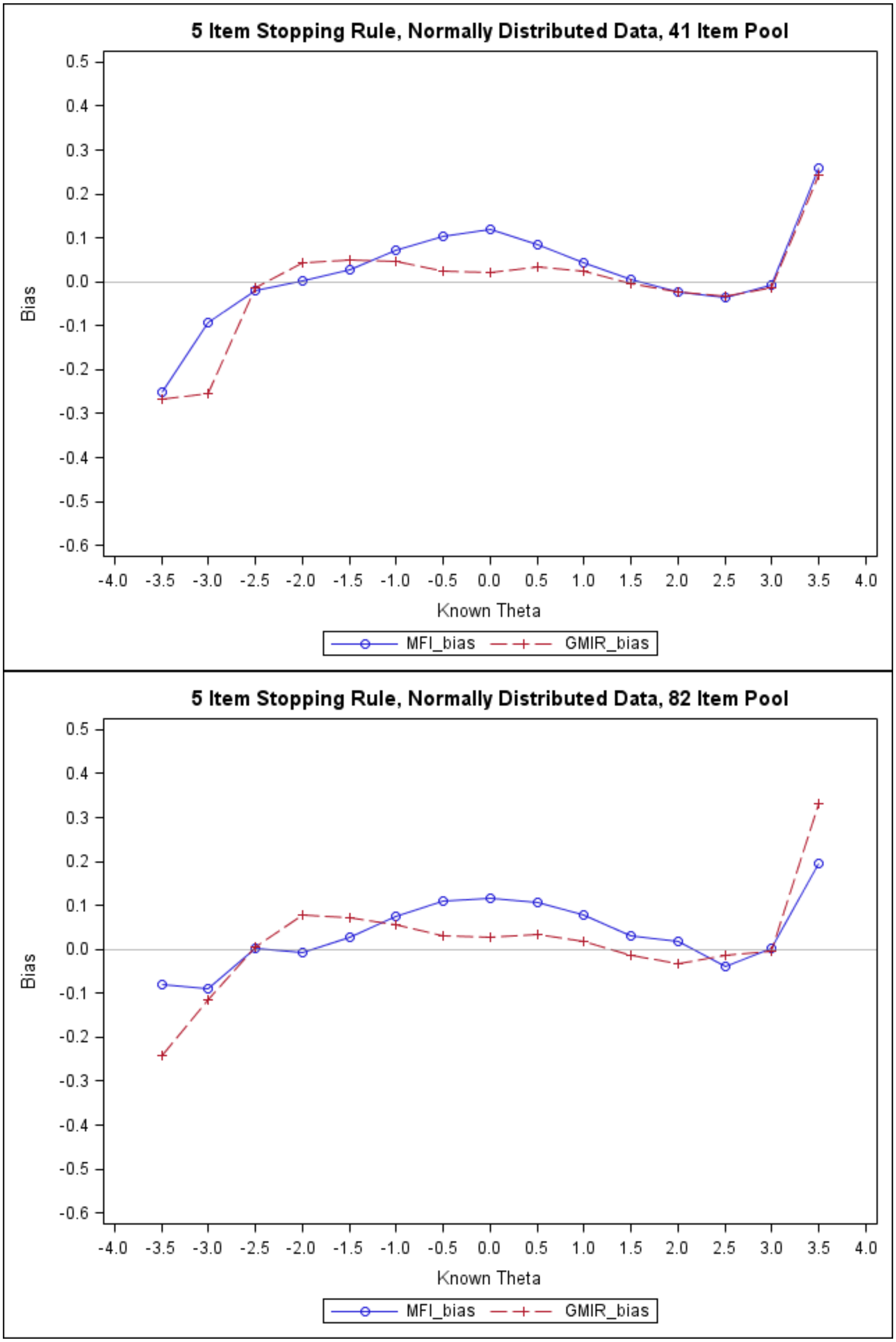
Table 5. Bias and RMSE Averaged across 100 Replications

CONDITIONAL MEASUREMENT PRECISION

Conditional plots of mean bias, mean RMSE, and grand mean SE averaged across the 100 samples show the precision of the final theta across a range of known θ values (-2, -1, 0, 1, 2) and the interim θ values across the items of the test. Plots of bias, RMSE, and SE conditional on known θ illustrate how MFI and GMIR performed in terms of measurement precision across the range of θ values, from -3.5 to 3.5. Bias values conditional on known thetas for each of the 12 simulated conditions are shown in Figures 5 and 6. For most of the θ scale, MFI and GMIR resulted in similar bias values. However, at two points on the θ scale one item selection method outperformed the other. At θ=-3.5, the bias was larger in almost all of the GMIR conditions. Around θ=0, the bias was larger in the MFI, 5 item stopping rule conditions. There were not patterns across or among item pool or population distribution conditions.

RMSE values displayed a generally consistent pattern across θ values across conditions, with RMSE values higher in GMIR conditions at the negative end of the θ scale and higher in MFI conditions in the middle of the θ scale, as shown in Figures 7 and 8. GMIR generally resulted in larger RMSE for θs less than -1. This pattern across θ values could be explained by the information function of the item pool. The item pool provides greater information in the middle of the θ values and less information at either end of the scale. Two conditions, the 5 and 9 item conditions with a normally distributed population and a 41 item pool did show a larger RMSE using MFI at θ=-3.5. Between θs of -1 and 1.5, MFI generally resulted in larger RMSE, with 5 item stopping rule conditions having the largest differences. When θ was greater than 1.5, GMIR generally resulted in a similar or larger RMSE. In the 7 item/normal/82 pool and 9 item/normal/ 41 pool at θ=3.5, GMIR resulted in a larger RMSE.

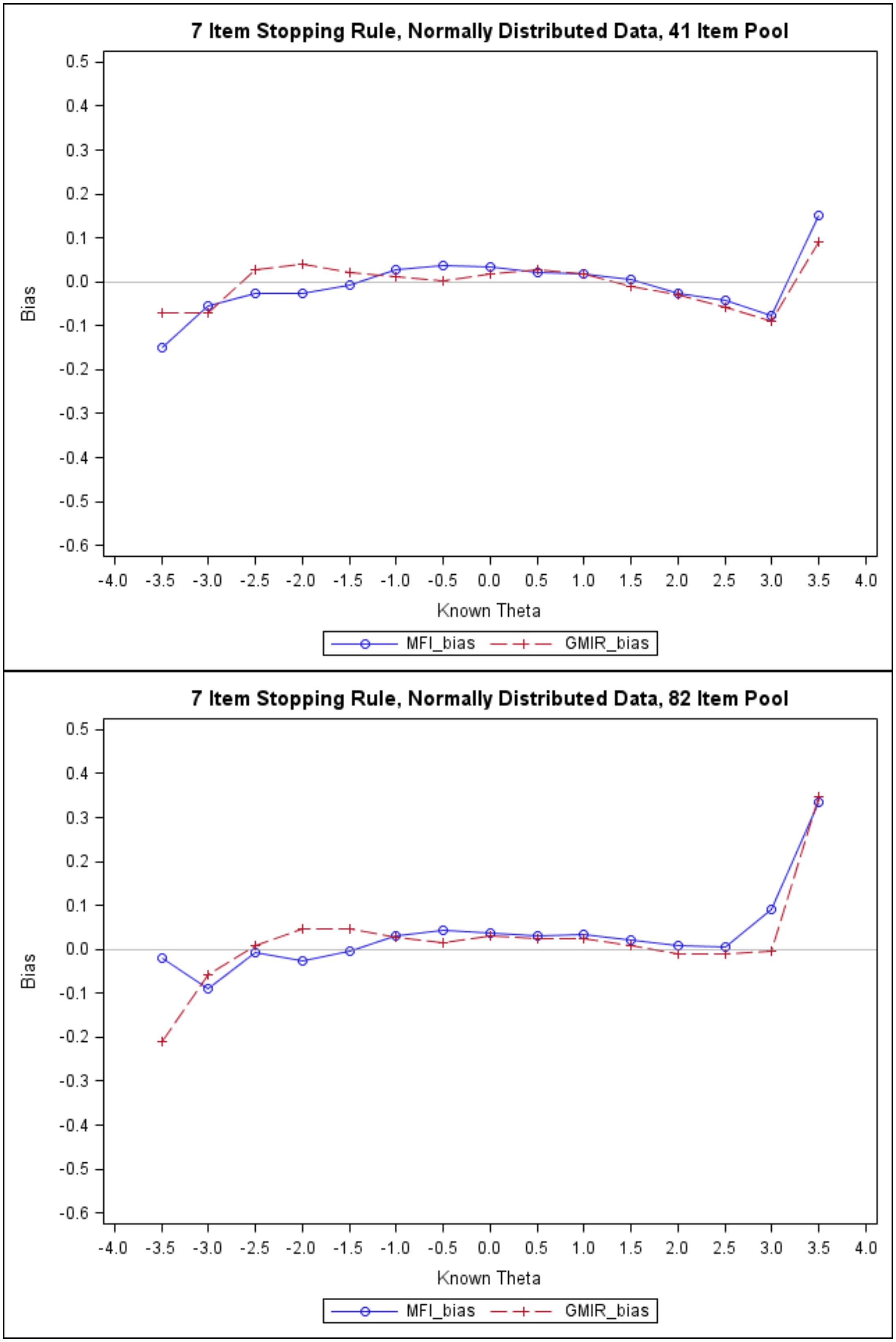
The SE conditional on the known θ also displayed a consistent pattern across simulation conditions, with SE values higher in most GMIR conditions at either end of the θ scale and comparable SE values in the middle of the θ scale, as shown in Figures 9 and 10. When the θ was less than -1.5, GMIR resulted in a larger SE than MFI, with the difference getting larger as the θ was smaller. The difference was greater when the 82 item pool and 5 or 9 item stopping rule was used. When the normally distributed population was used and there were simulees with θ values at the upper end of the scale, GMIR resulted in a larger SE than MFI when θ was greater than 2.0/2.5 for half of the conditions: in the 5 item stopping rule/82 item pool, 7 item stopping rule/82 item pool, and 9 item stopping rule/41 item pool conditions. For the other normal conditions, the MFI and GMIR values were similar. In the middle of the scale, MFI and GMIR resulted in similar SE values.



5 Item Stopping Rule, Normally Distributed Data, 41 Item Pool

5 Item Stopping Rule, Normally Distributed Data, 82 Item Pool

Figure 5A. Plots of Mean Bias Conditional on Known Theta for the 5 Item Stopping Rule and Normally Distributed Population Conditions


7 Item Stopping Rule, Normally Distributed Data, 41 Item Pool

7 Item Stopping Rule, Normally Distributed Data, 82 Item Pool

Figure 5B. Plots of Mean Bias Conditional on Known Theta for the 7 Item Stopping Rule and Normally Distributed Population Conditions

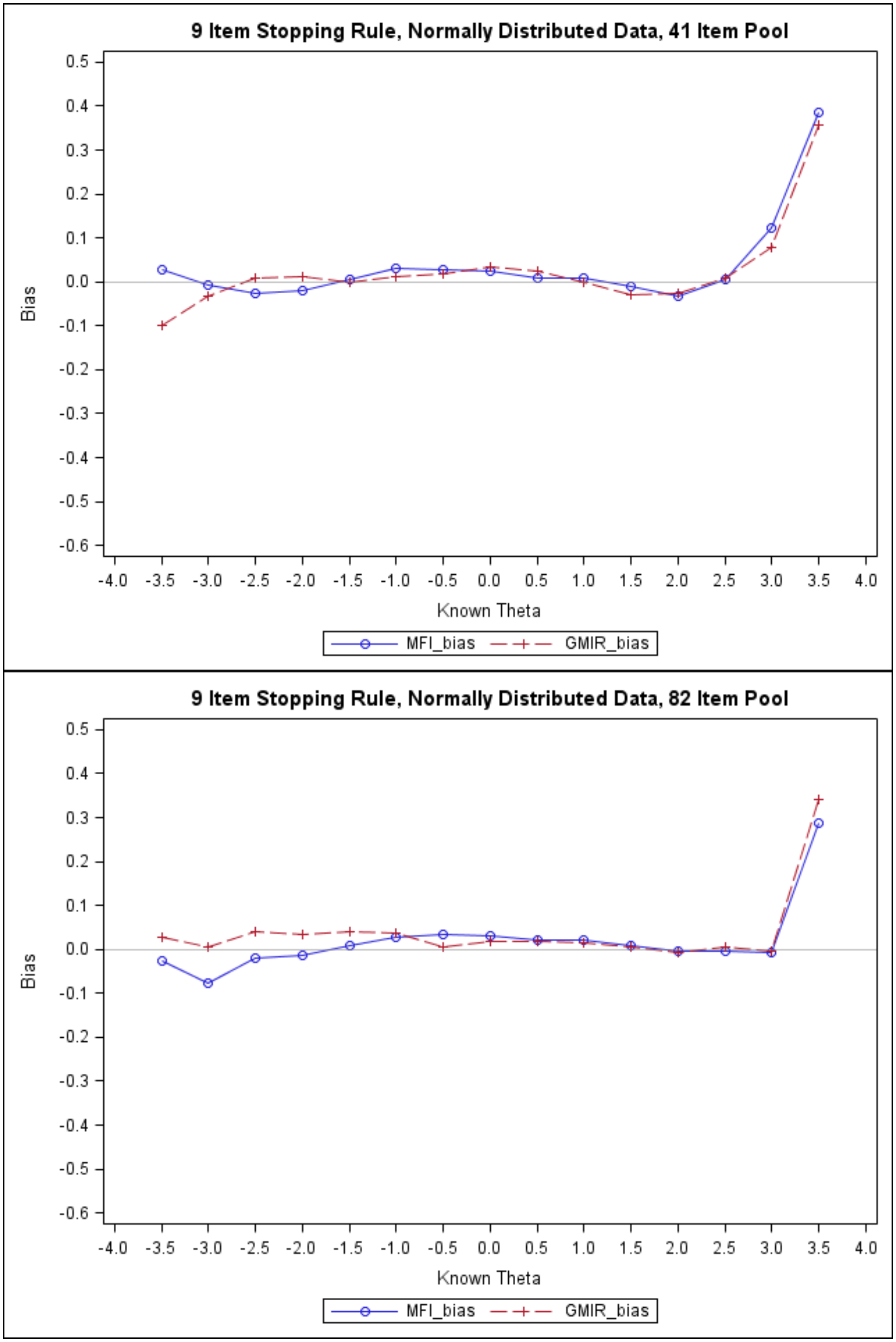
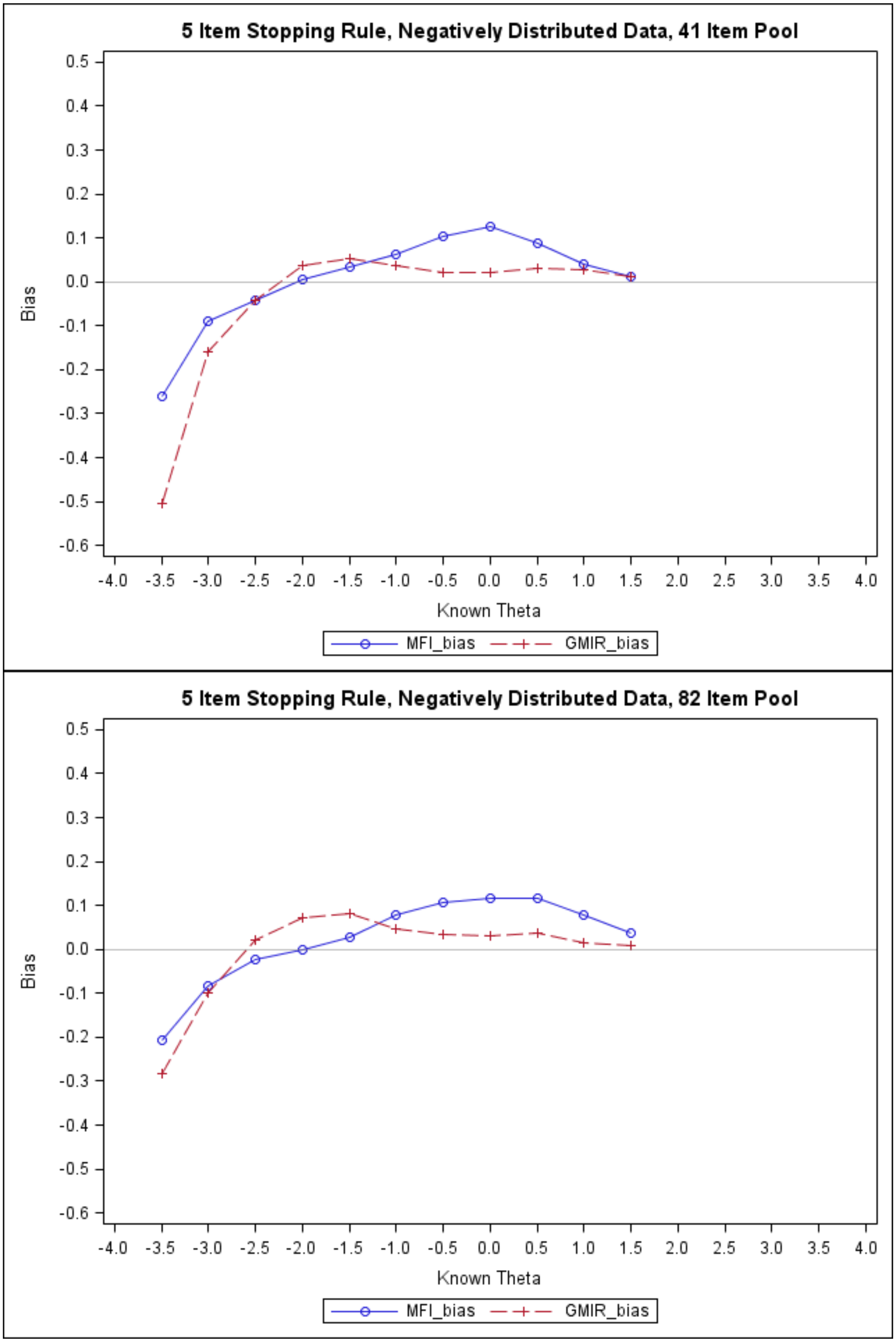


Figure 5C. Plots of Mean Bias Conditional on Known Theta for the 9 Item Stopping Rule and Normally Distributed Population Conditions



5 Item Stopping Rule, Negatively Distributed Data, 41 Item Pool

5 Item Stopping Rule, Negatively Distributed Data, 82 Item Pool

Figure 6A. Plots of Mean Bias Conditional on Known Theta for the 5 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

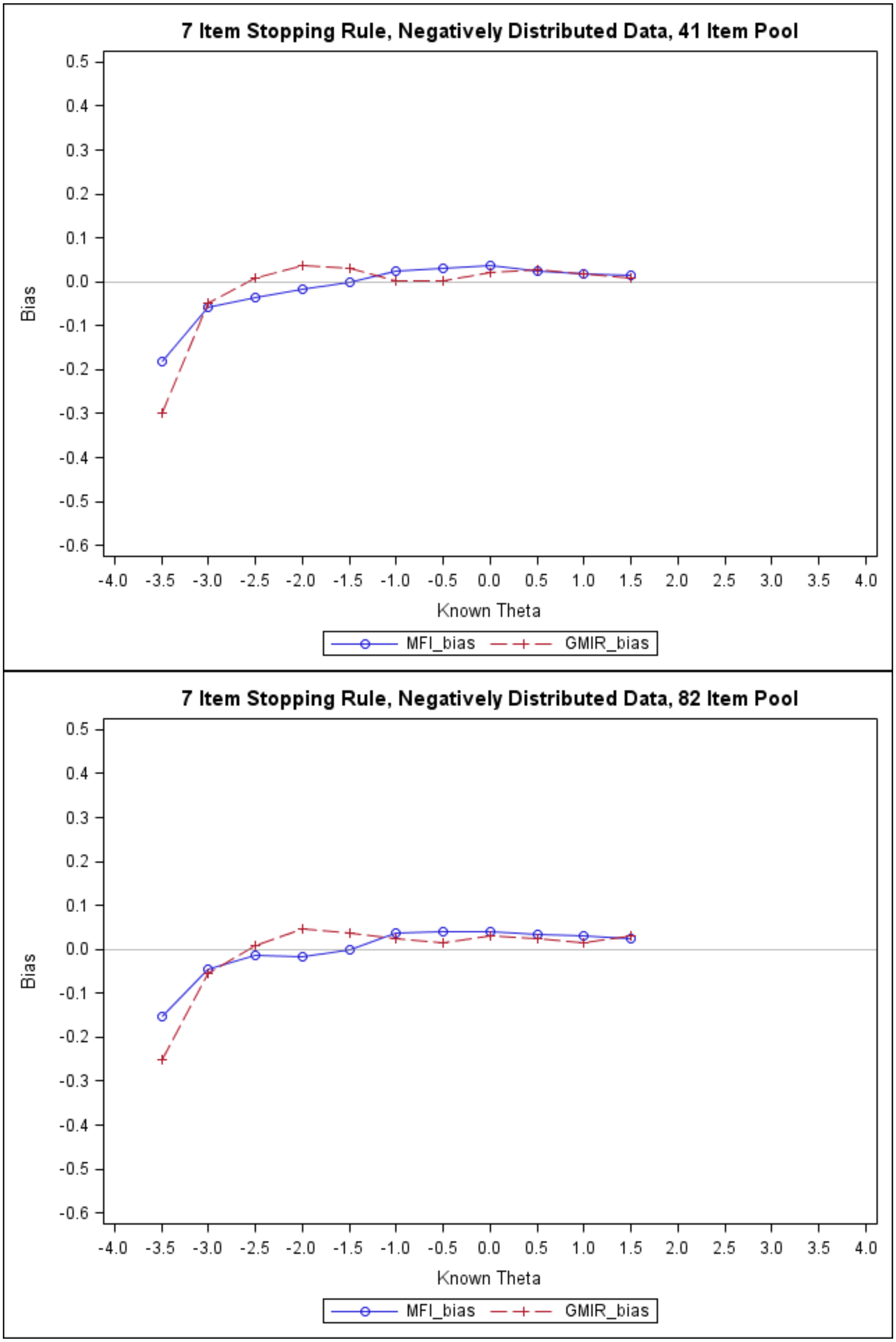
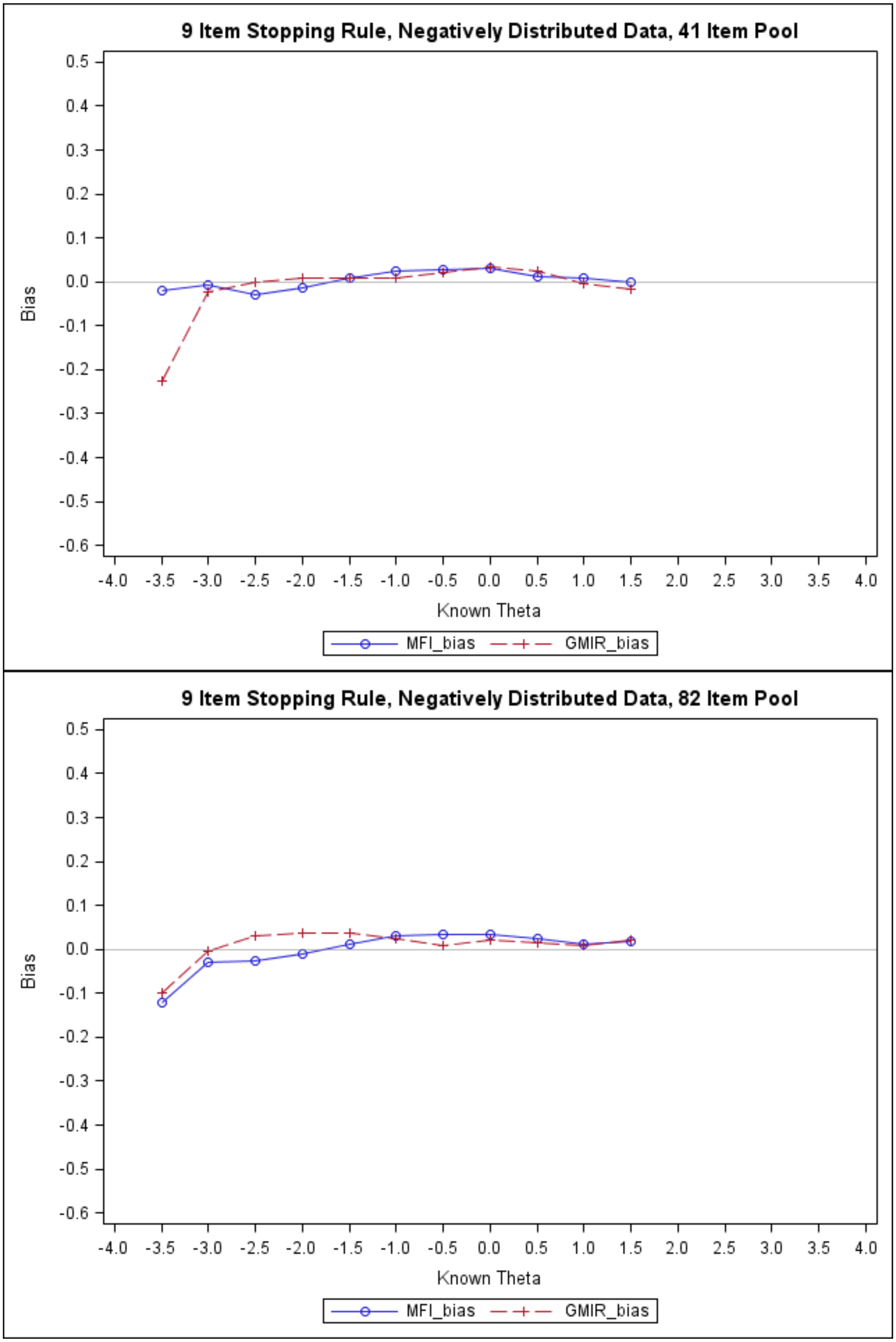


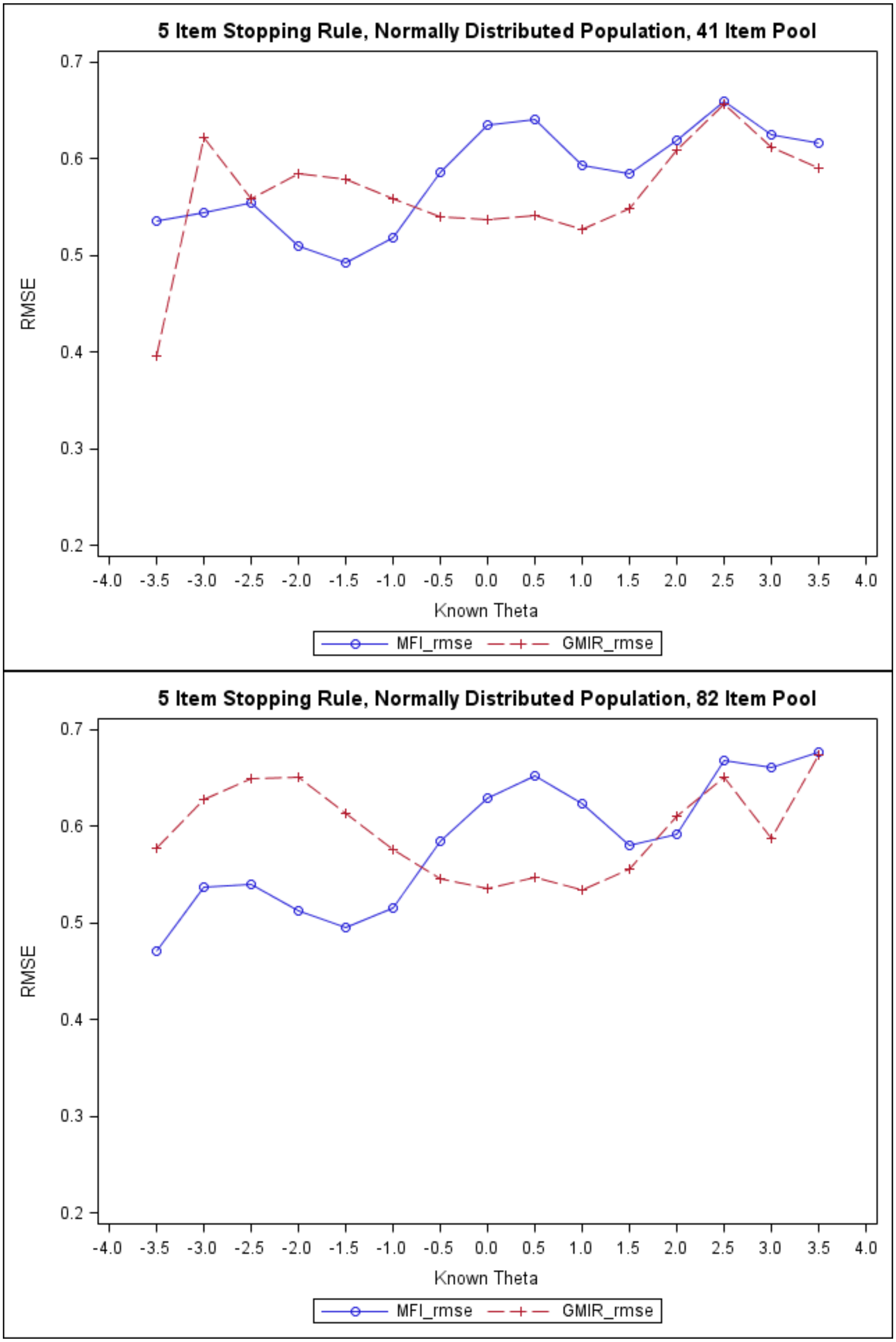
Figure 6B. Plots of Mean Bias Conditional on Known Theta for the 7 Item Stopping Rule and Negatively Skewed Distributed Population Conditions



9 Item Stopping Rule, Negatively Distributed Data, 41 Item Pool

9 Item Stopping Rule, Negatively Distributed Data, 82 Item Pool

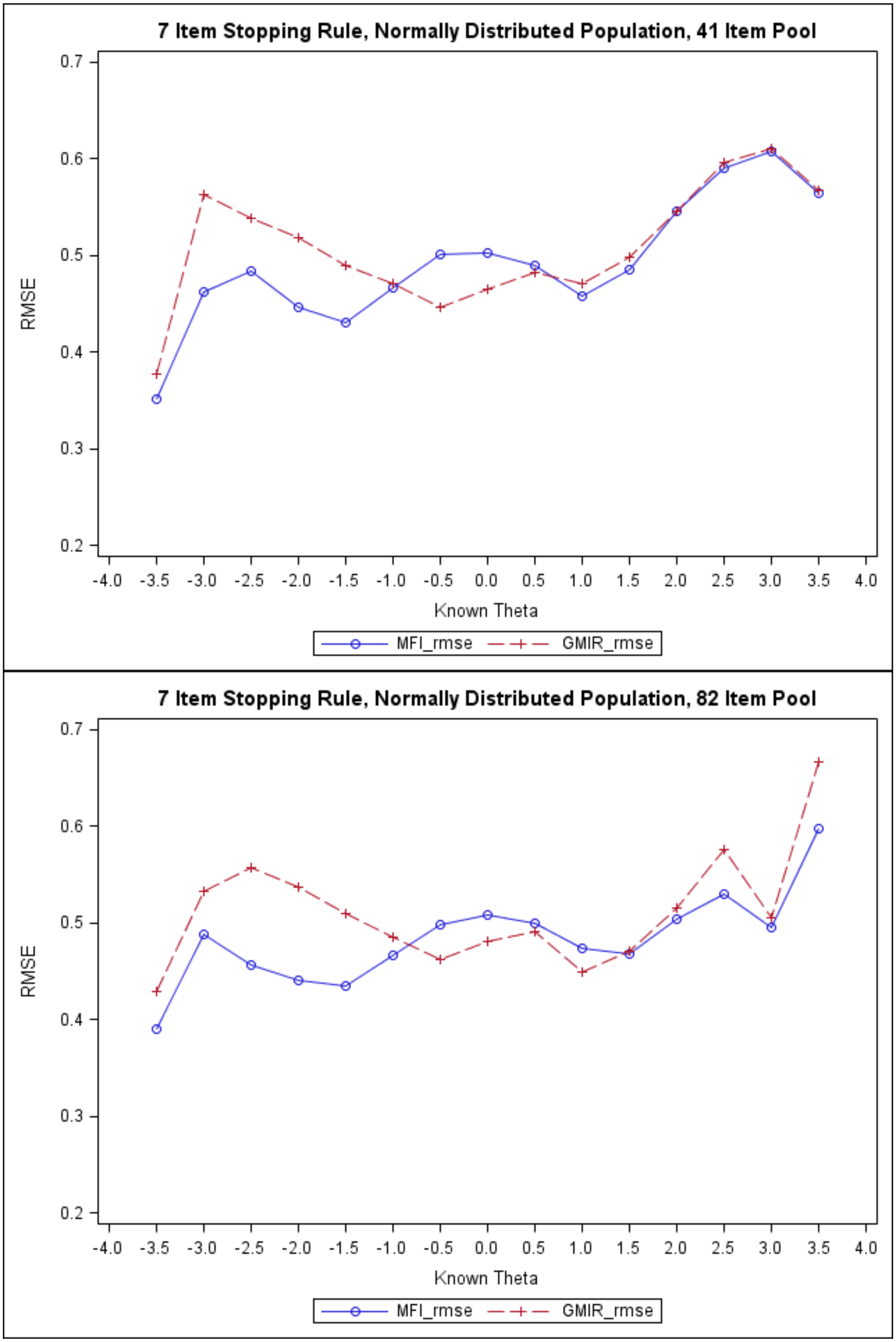
Figure 6C. Plots of Mean Bias Conditional on Known Theta for the 9 Item Stopping Rule and Negatively Skewed Distributed Population Conditions



5 Item Stopping Rule, Normally Distributed Population, 41 Item Pool

5 Item Stopping Rule, Normally Distributed Population, 82 Item Pool

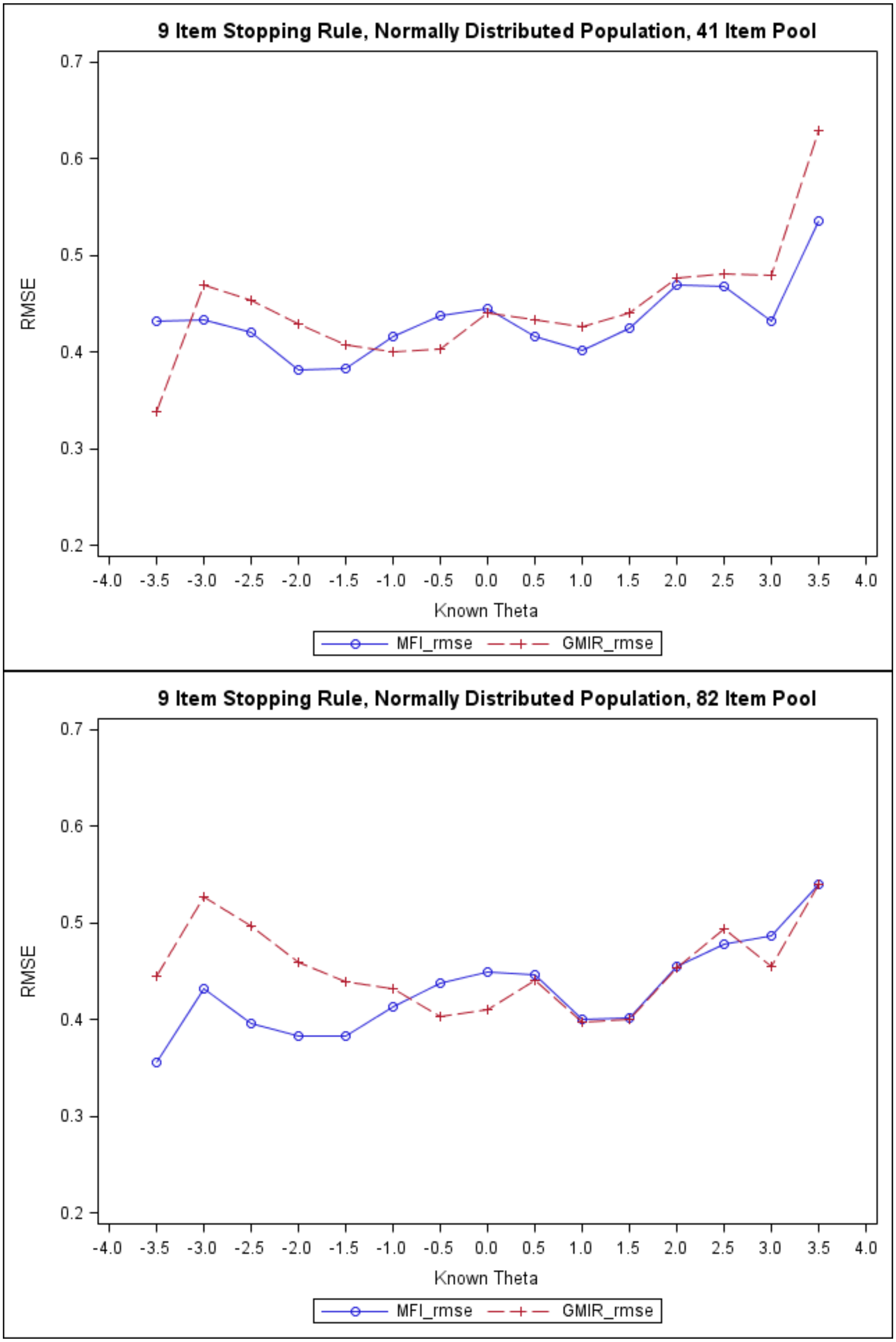
Figure 7A. Plots of Mean RMSE Conditional on Known Theta for the 5 Item Stopping Rule and Normally Distributed Population Conditions



7 Item Stopping Rule, Normally Distributed Population, 41 Item Pool

7 Item Stopping Rule, Normally Distributed Population, 82 Item Pool

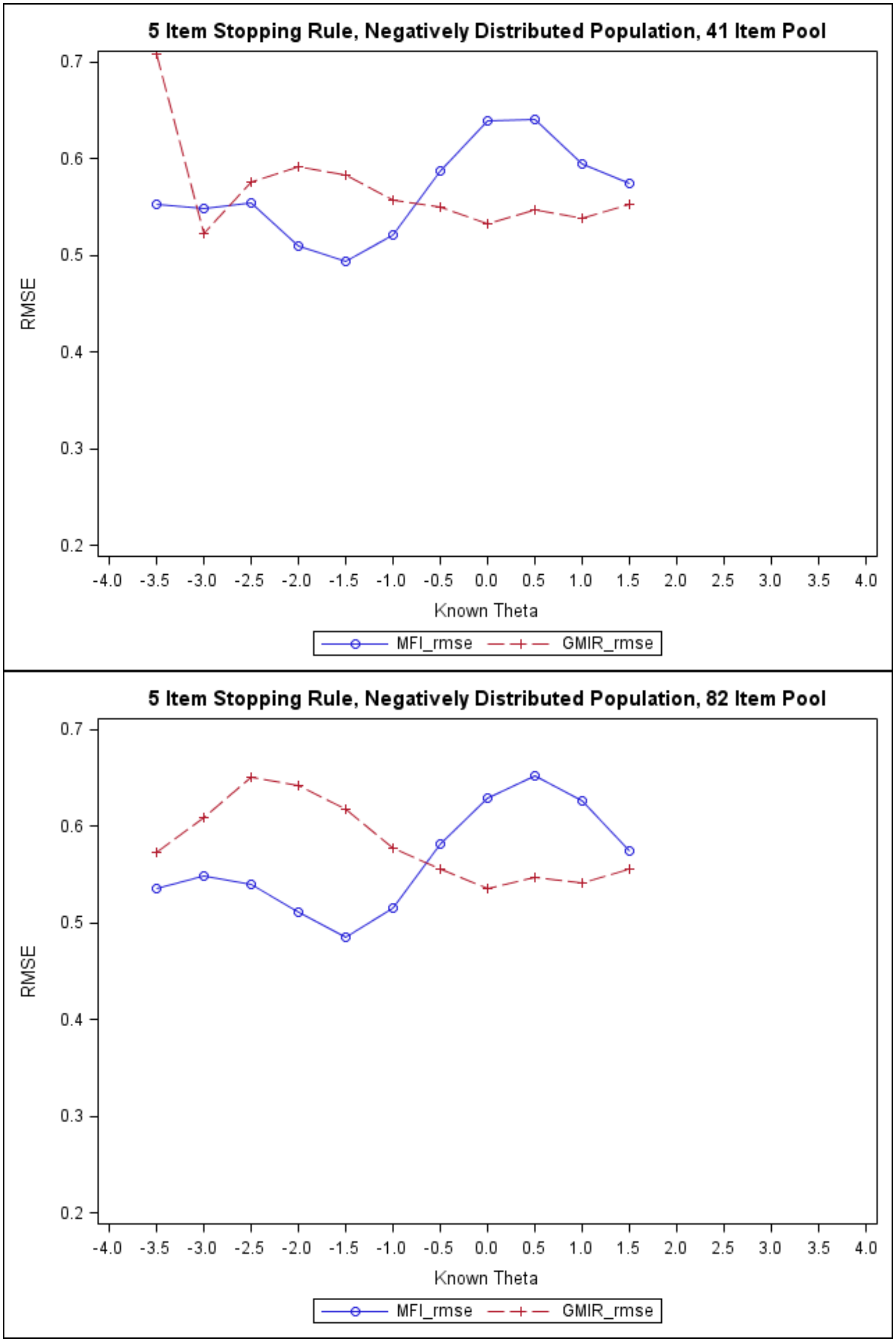
Figure 7B. Plots of Mean RMSE Conditional on Known Theta for the 7 Item Stopping Rule and Normally Distributed Population Conditions



9 Item Stopping Rule, Normally Distributed Population, 41 Item Pool

9 Item Stopping Rule, Normally Distributed Population, 82 Item Pool

Figure 7C. Plots of Mean RMSE Conditional on Known Theta for the 9 Item Stopping Rule and Normally Distributed Population Conditions



5 Item Stopping Rule, Negatively Distributed Population, 41 Item Pool

5 Item Stopping Rule, Negatively Distributed Population, 82 Item Pool

Figure 8A. Plots of Mean RMSE Conditional on Known Theta for the 5 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

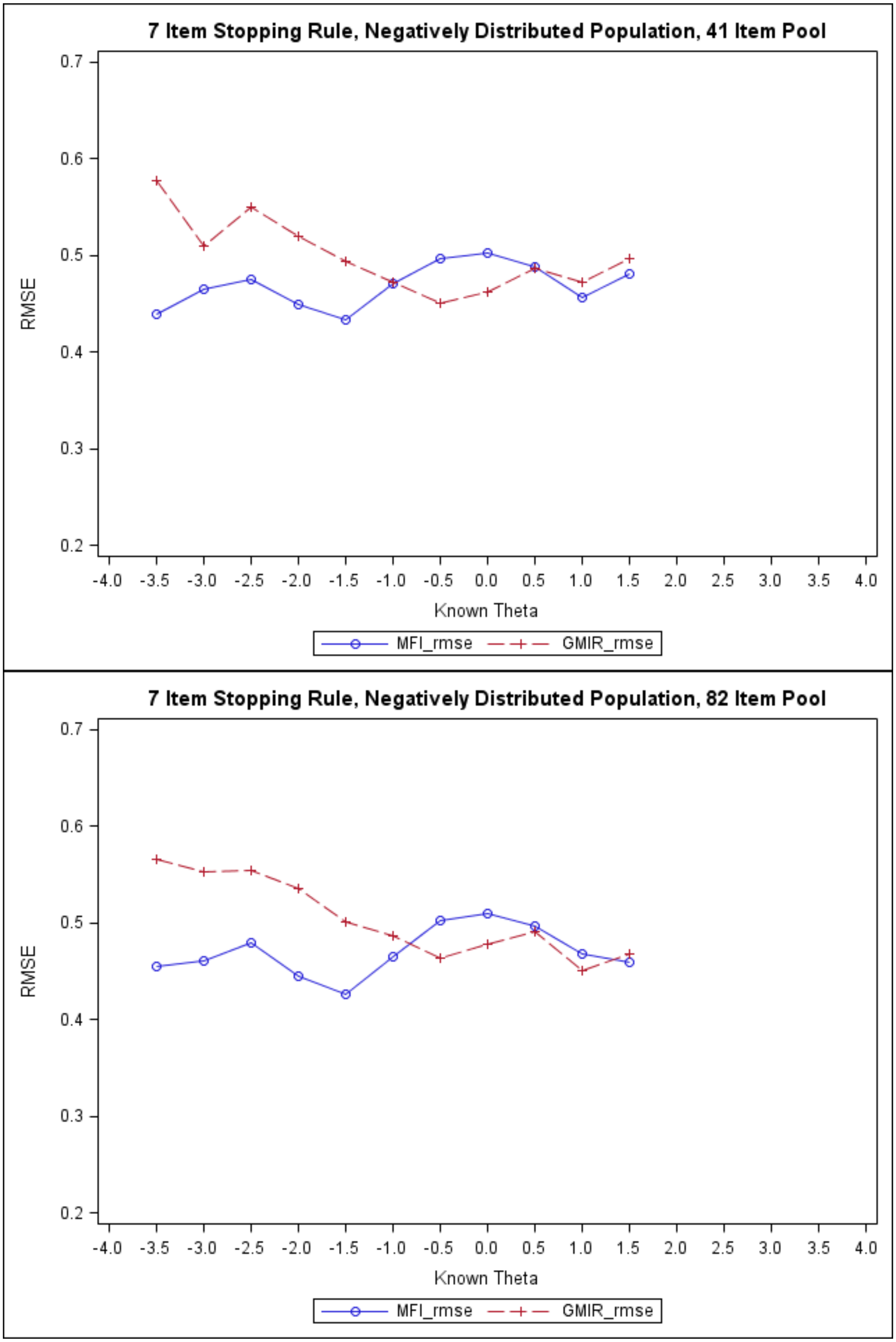


Figure 8B. Plots of Mean RMSE Conditional on Known Theta for the 7 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

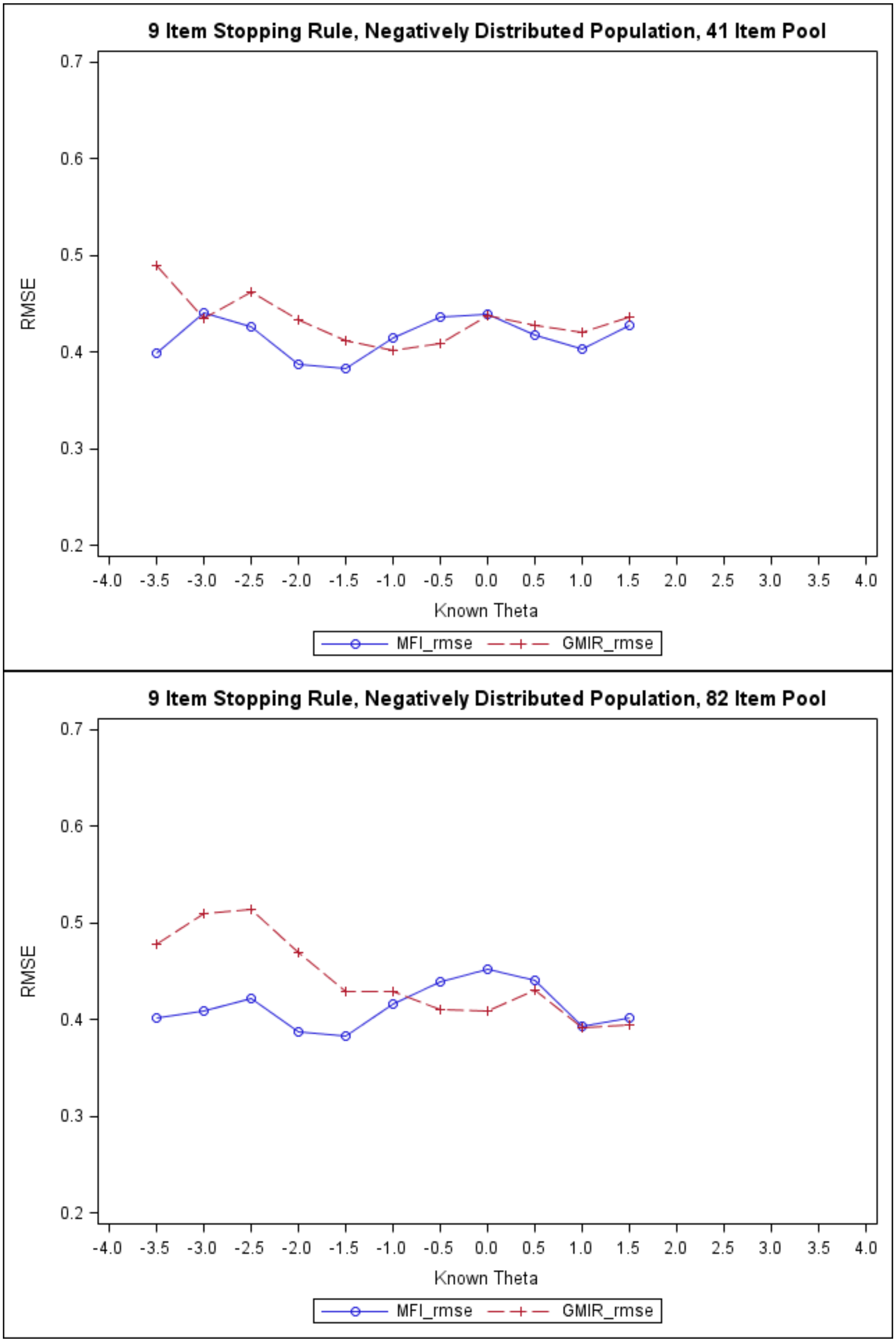


Figure 8C. Plots of Mean RMSE Conditional on Known Theta for the 9 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

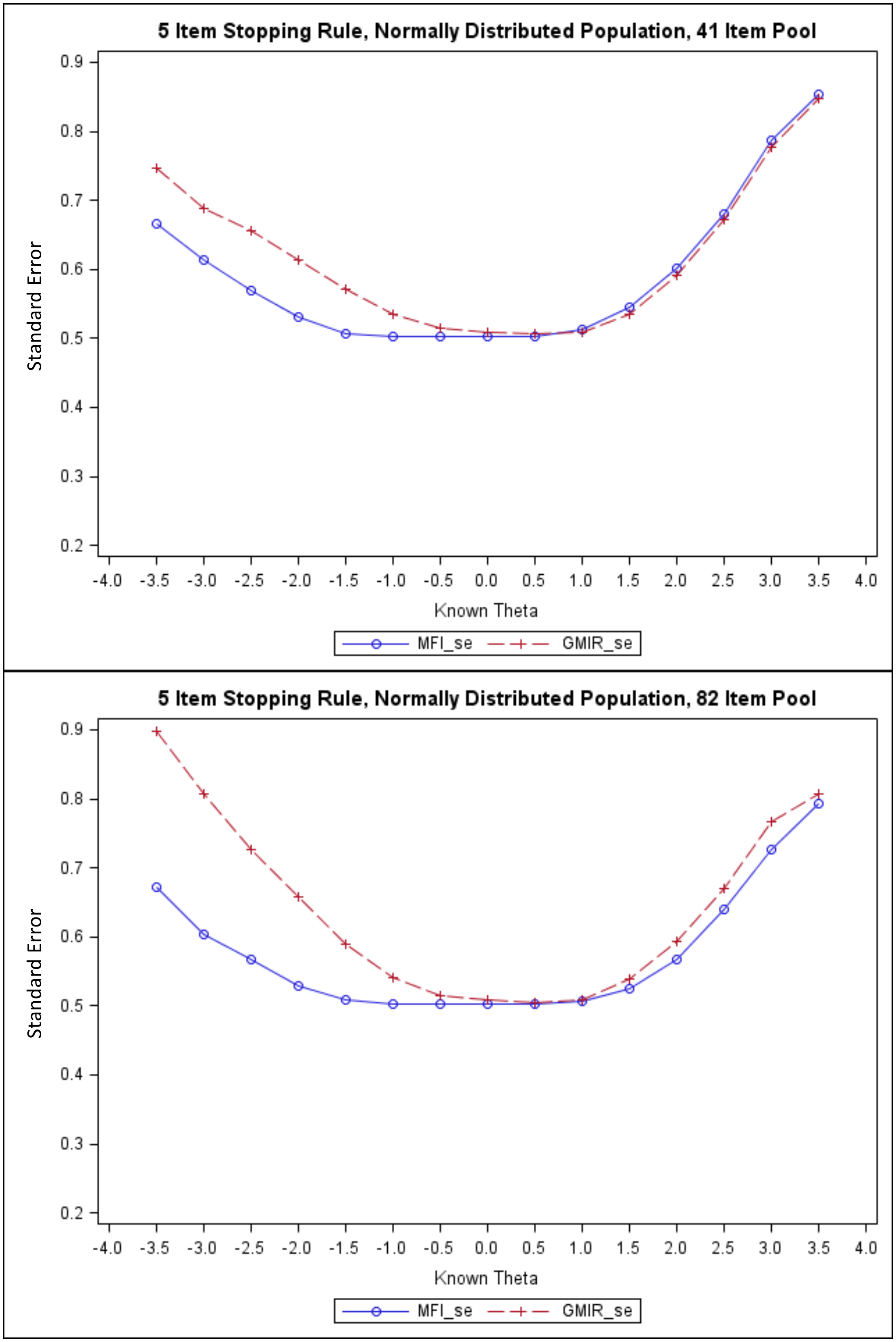


Figure 9A. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 5 Item Stopping Rule and Normally Distributed Population Conditions

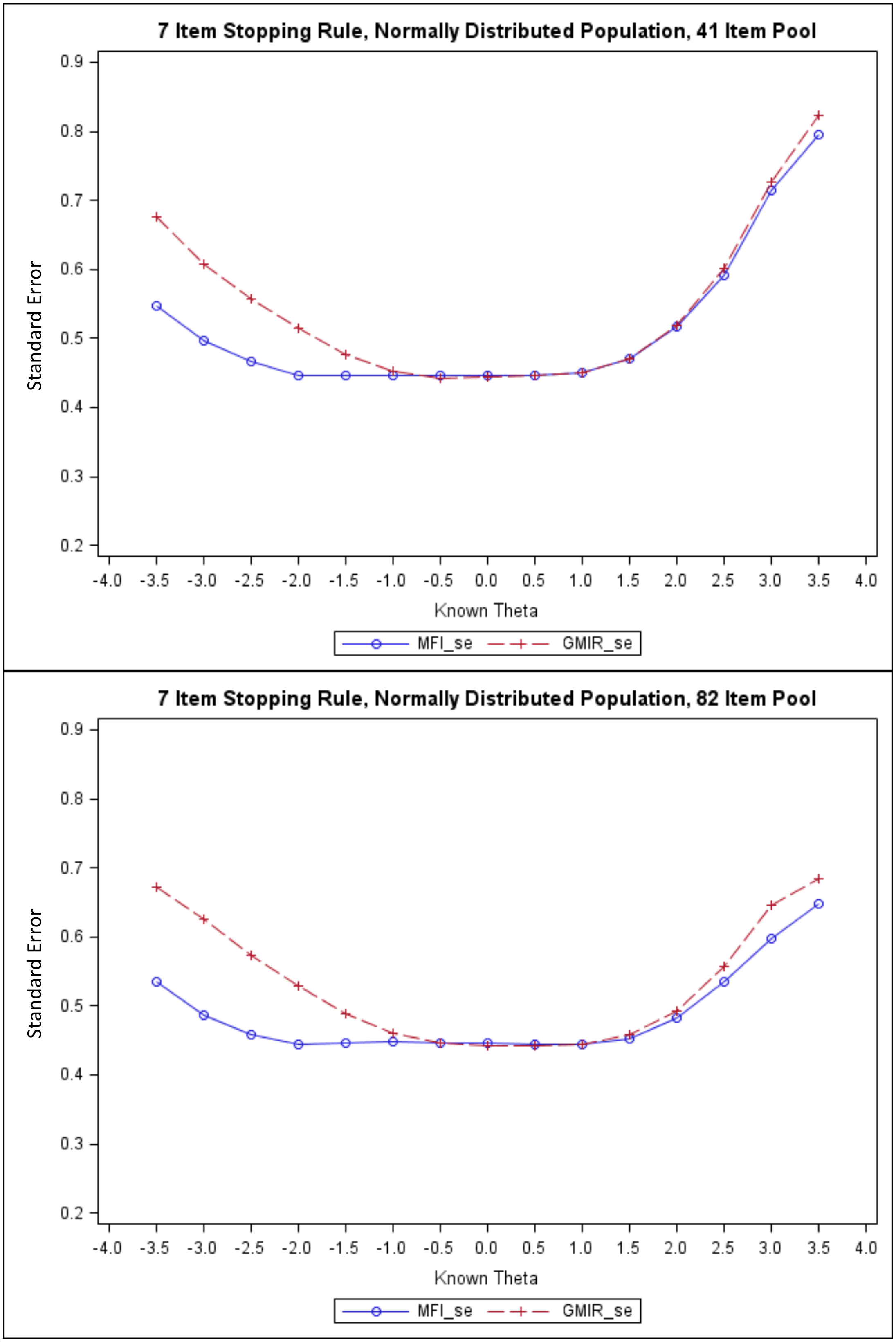
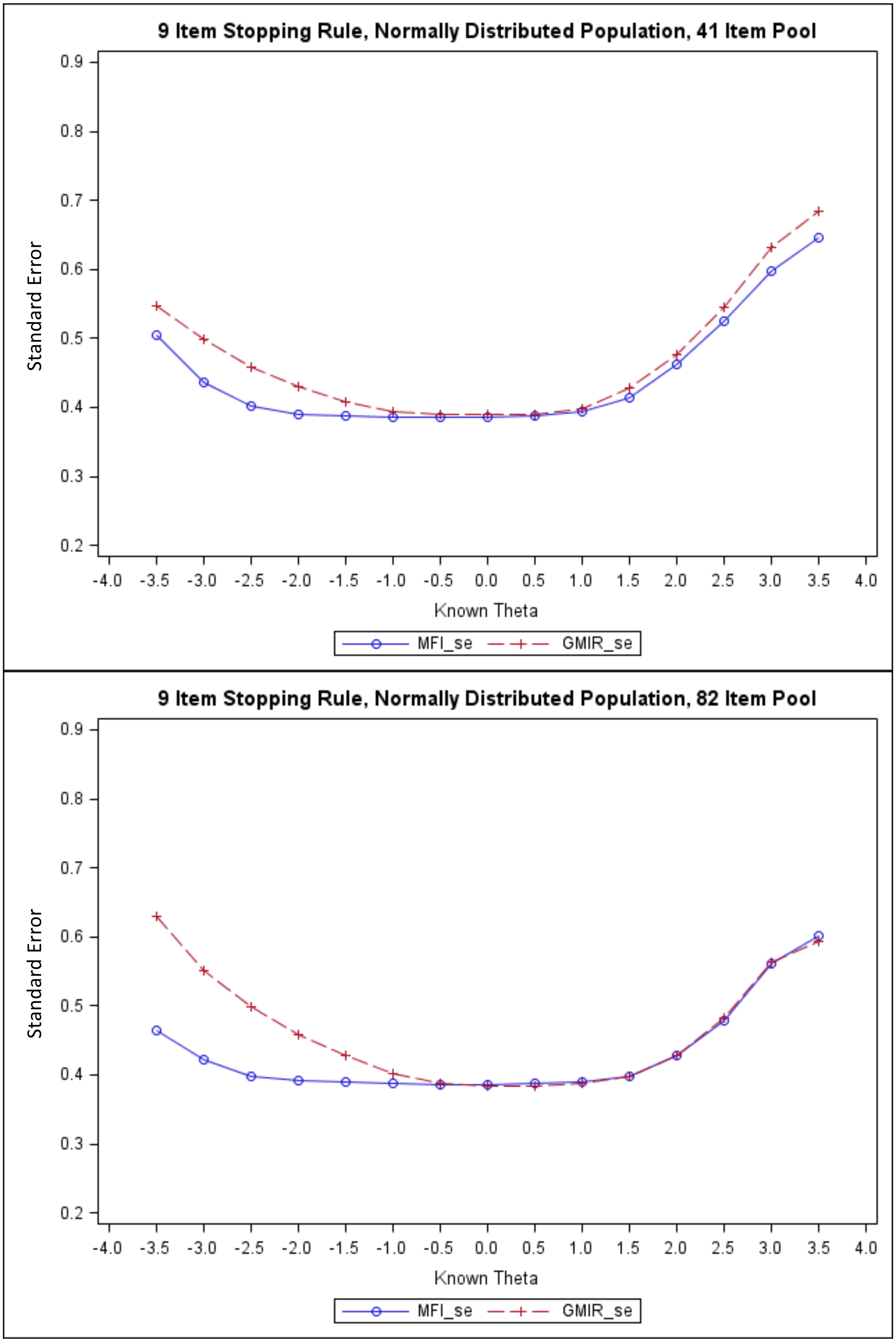


Figure 9B. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 7 Item Stopping Rule and Normally Distributed Population Conditions



9 Item Stopping Rule, Normally Distributed Population, 41 Item Pool

9 Item Stopping Rule, Normally Distributed Population, 82 Item Pool

Figure 9C. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 9 Item Stopping Rule and Normally Distributed Population Conditions

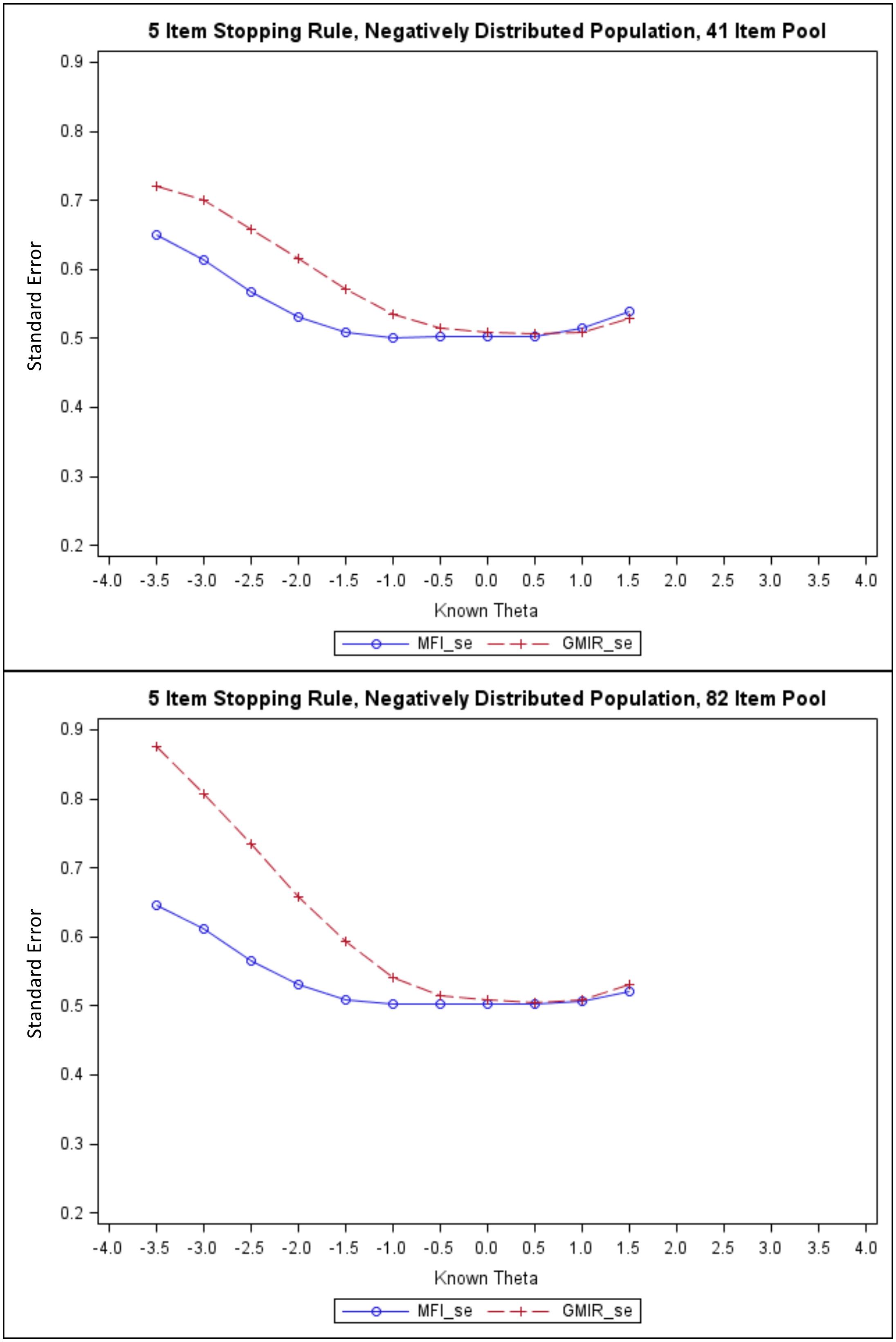


Figure 10A. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 5 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

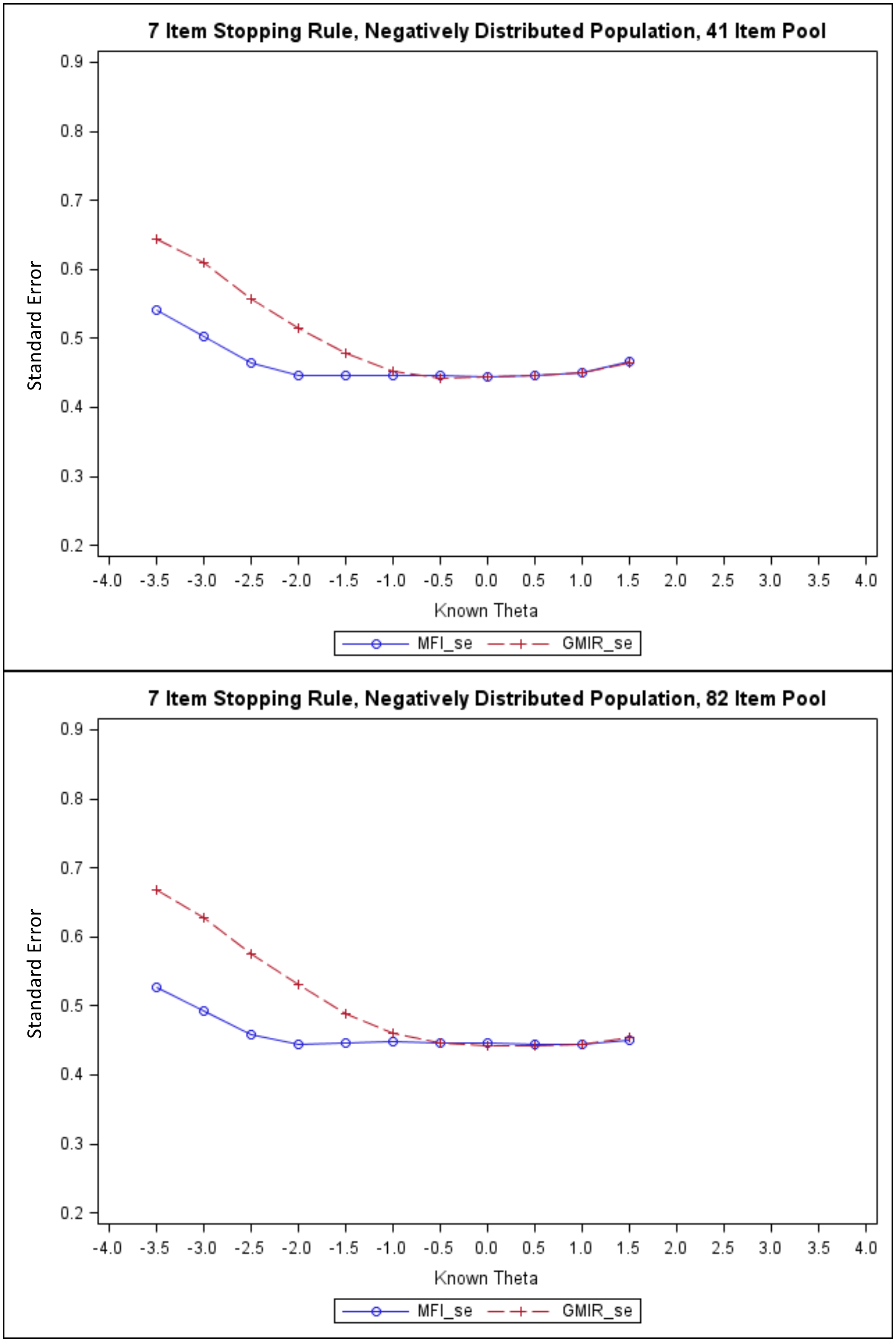
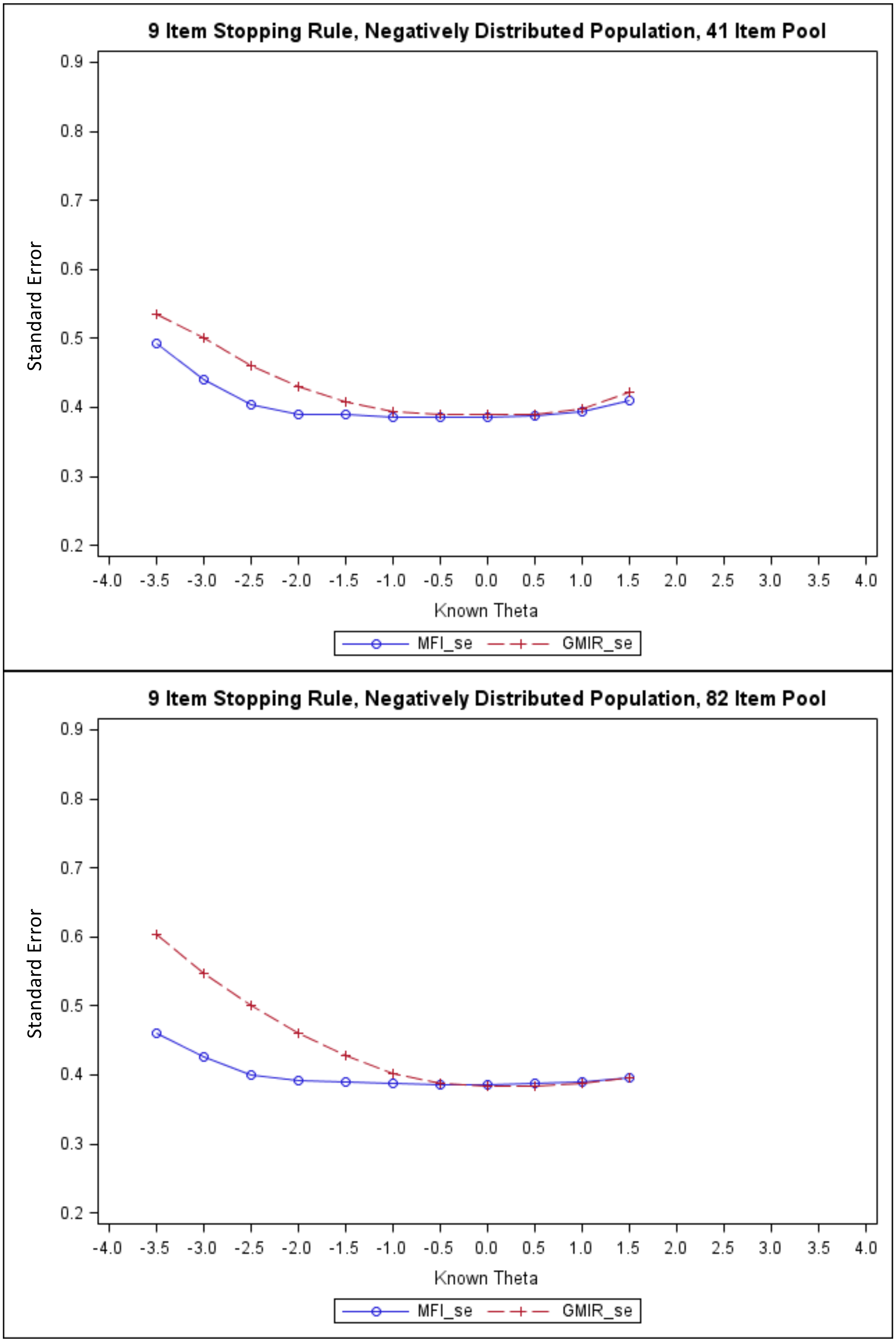


Figure 10B. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 7 Item Stopping Rule and Negatively Skewed Distributed Population Conditions



9 Item Stopping Rule, Negatively Distributed Population, 41 Item Pool

9 Item Stopping Rule, Negatively Distributed Population, 82 Item Pool

Figure 10C. Plots of Mean Standard Error (SE) Conditional on Known Theta for the 9 Item Stopping Rule and Negatively Skewed Distributed Population Conditions

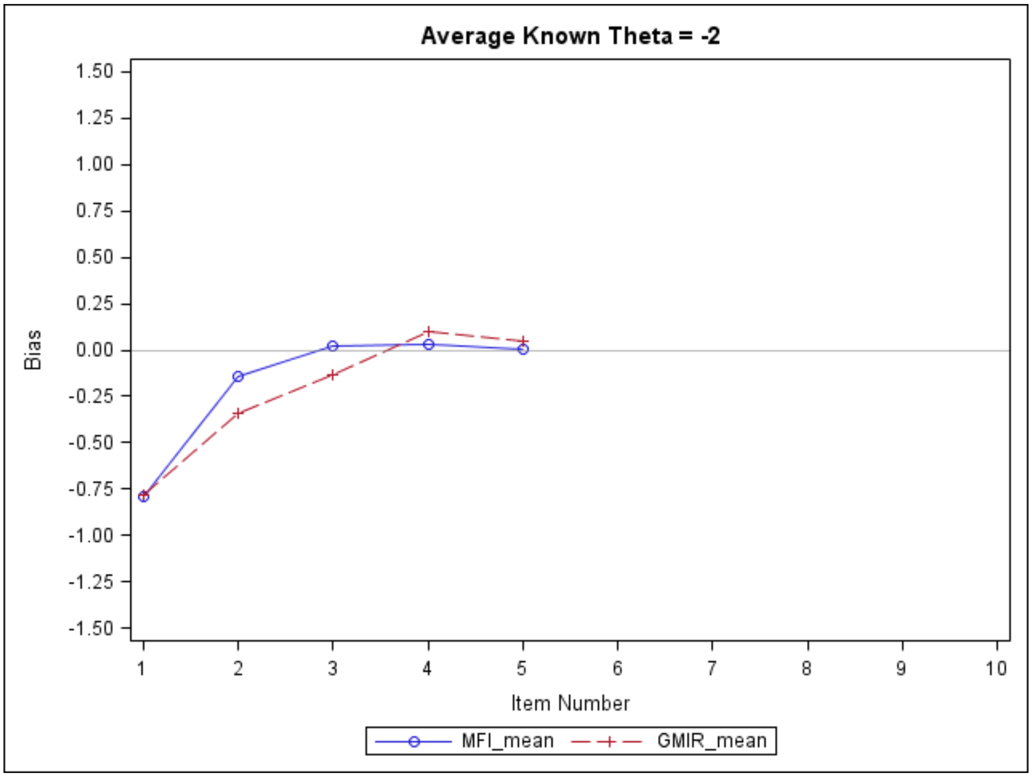
Plots of bias, RMSE, and SE conditional on the item number, illustrate how MFI and GMIR performed in terms of measurement precision at each item administered, averaged across the 100 samples. Examinees were divided into 5 groups according to their known θ: -2, -1, 0, 1, and 2. Conditional plots for each theta group for each condition are reported in Figures 11 through 26. Conditions using the stopping rule of 5 items or .54 SE, were plotted items 1 through 5. Conditions using the stopping rule of 7 items or .46 SE were plotted items 1 though 7. Conditions using the stopping rule of 9 items or .40 SE were plotted items 1 though 9.

The bias by item did not vary between the normally distributed population conditions and the negatively skewed population conditions for any theta value, so only the normal conditions are shown, the negatively skewed plots and any other plots not shown can be found in the Appendix A. In general, across theta values and conditions, GMIR resulted in a larger bias value at the beginning of the test (especially at the negative theta values) and MFI resulted in a larger bias at the end of the test (especially with a longer test). Conditional plots of bias by item number for theta =-2 are shown in Figures 11. Across all conditions for examinees in the group of known, MFI resulted in less bias at items 2 and 3. When the 5 item stopping rule was used, GMIR and MFI resulted in similar bias at the other items. When the 7 or item stopping rules were used, GMIR resulted in less bias at items 7 or 9 respectively. If the examinees were given all 7 or 9 items, this would be the final item administered. Some examinees were only administered 6 or 8 items and therefore would not be included in this average. This difference between GMIR and

MFI was larger at item 9 than at item 7. Conditional plots of bias by item number for theta=-1 are shown in Figures 12. Using the 5 item stopping rule, the bias by item plots are similar for MFI and GMIR. When the 7 and 9 item stopping rules were used, GMIR resulted in less bias at items 7 and 9 respectively. This difference was greater when the 41 item pool was used than the 82 item pool. MFI and GMIR resulted in similar bias at the other items.

Figures 13 show the conditional plots of bias by item number for theta=0. MFI resulted in less bias at items 2 and 3 across all conditions, the largest difference being at item 2. When the 7 item stopping rule was used, MFI resulted in a slightly smaller bias than GMIR at item 7. Figures 14 show the conditional plots of bias by item number for theta=1. Generally, across the items, MFI and GMIR resulted in similar bias values. MFI bias was slightly smaller at item 2 across conditions and also slightly smaller at item 1 when the 41 item pool was used. GMIR resulted in slightly less bias at item 9 in the 9 item stopping rule/41 item pool conditions. Conditional plots of bias by item number for theta=2 are shown in Figure 15. The pattern of bias across items for theta=2 was consistent across stopping rule, item pool, and population distribution conditions. Also, the item selection procedures MFI and GMIR resulted in almost identical in bias value patterns in each condition. Since the plots of all 12 conditions are similar, only the 9 item stopping rule/ normally distributed population/ 82 item pool condition is shown in the figure.

41 Item Pool



82 Item Pool

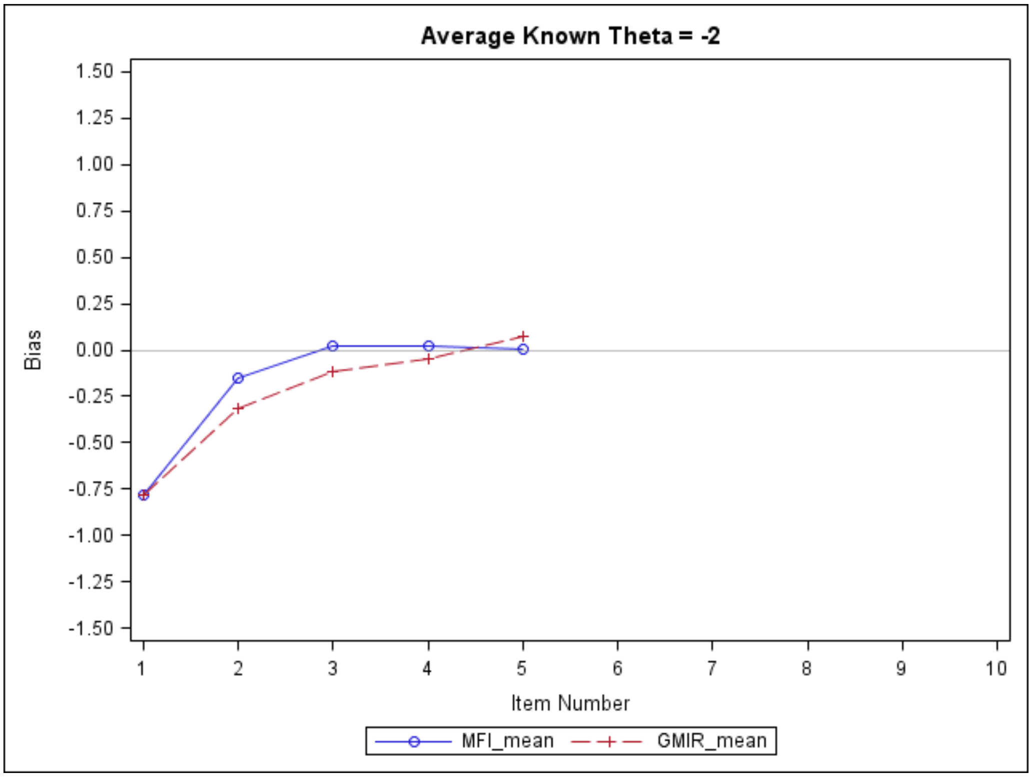
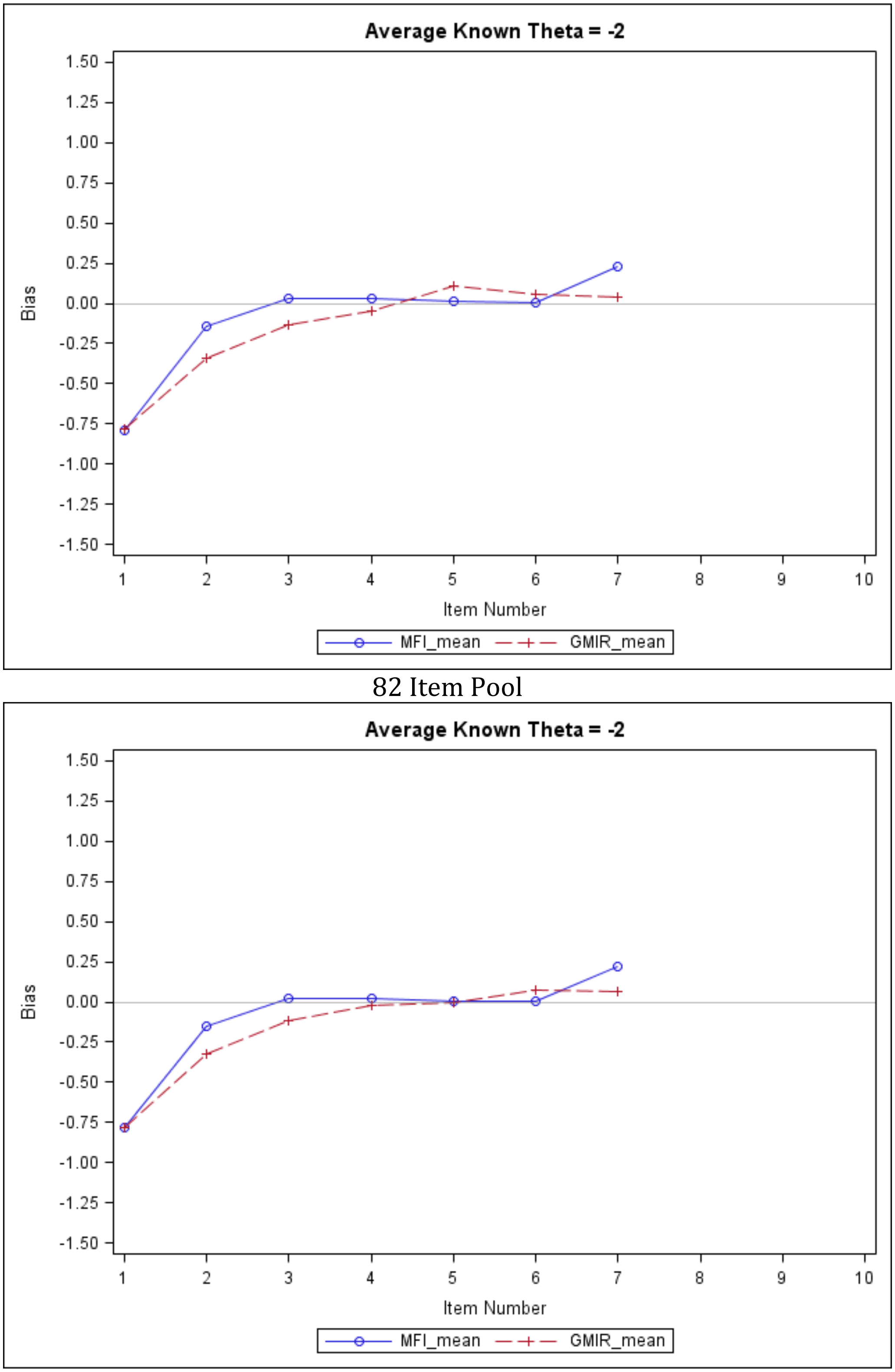
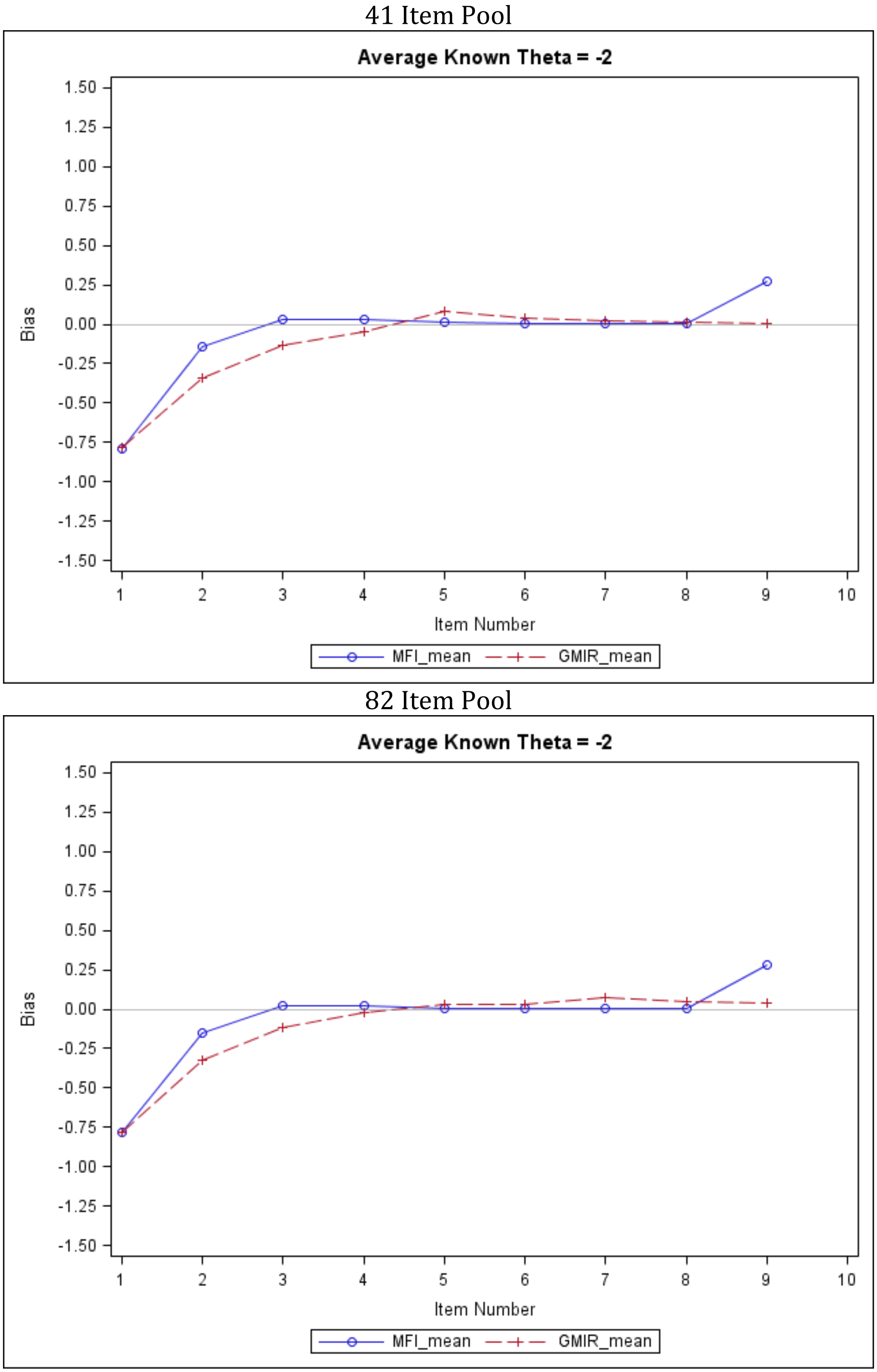


Figure 11A. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 11B. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 11C. Plots of Mean Bias Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Populations

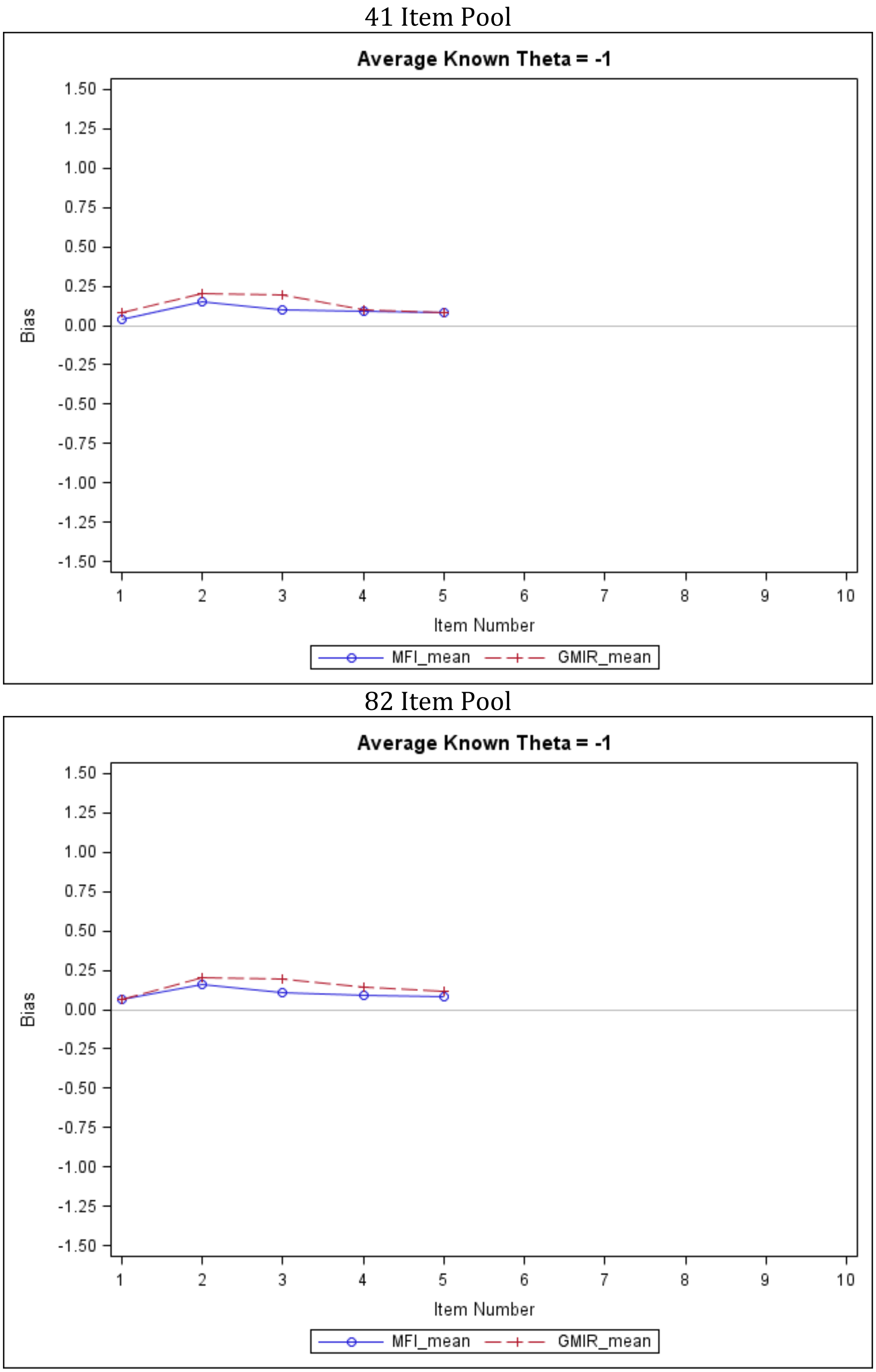
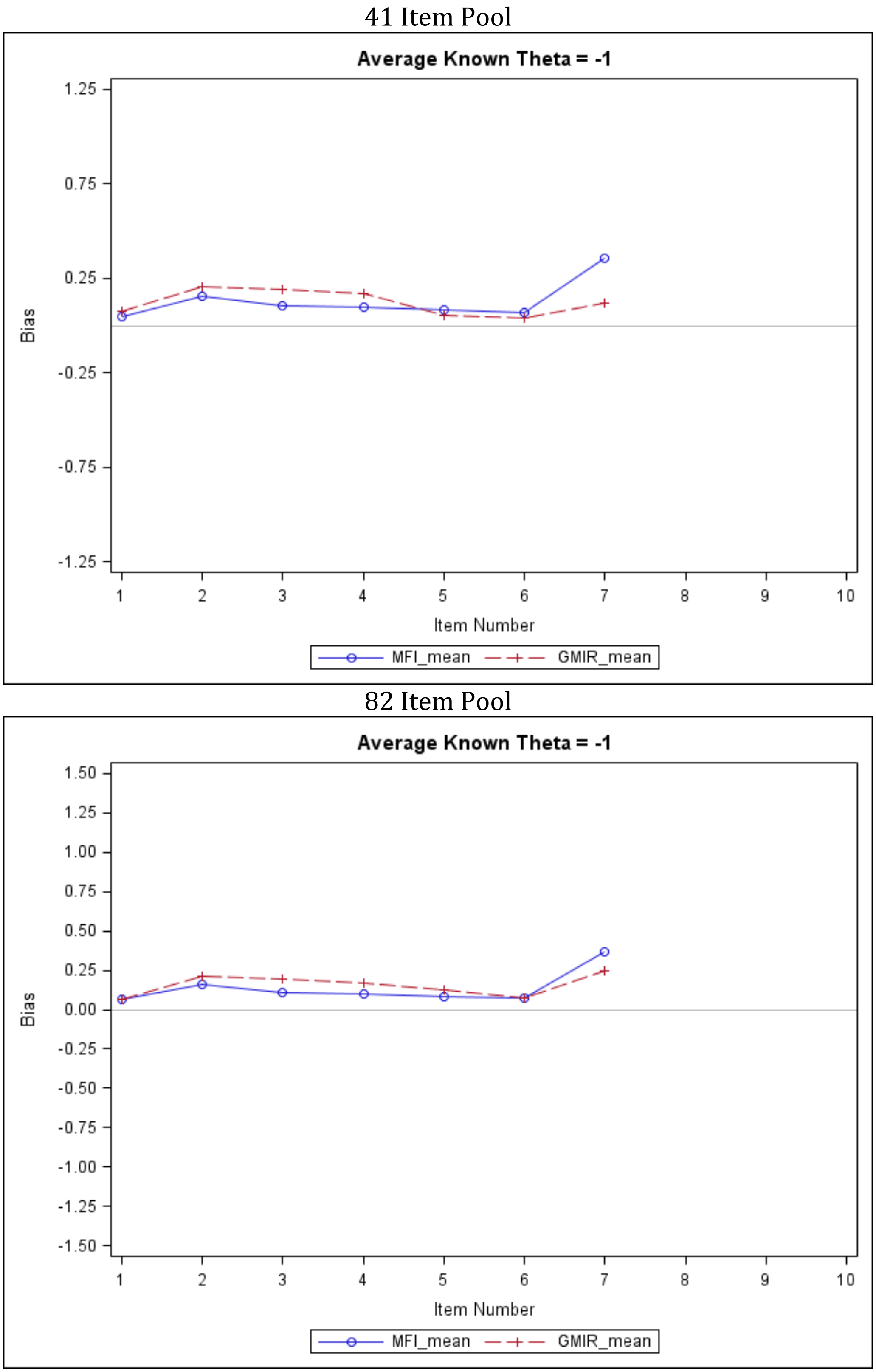


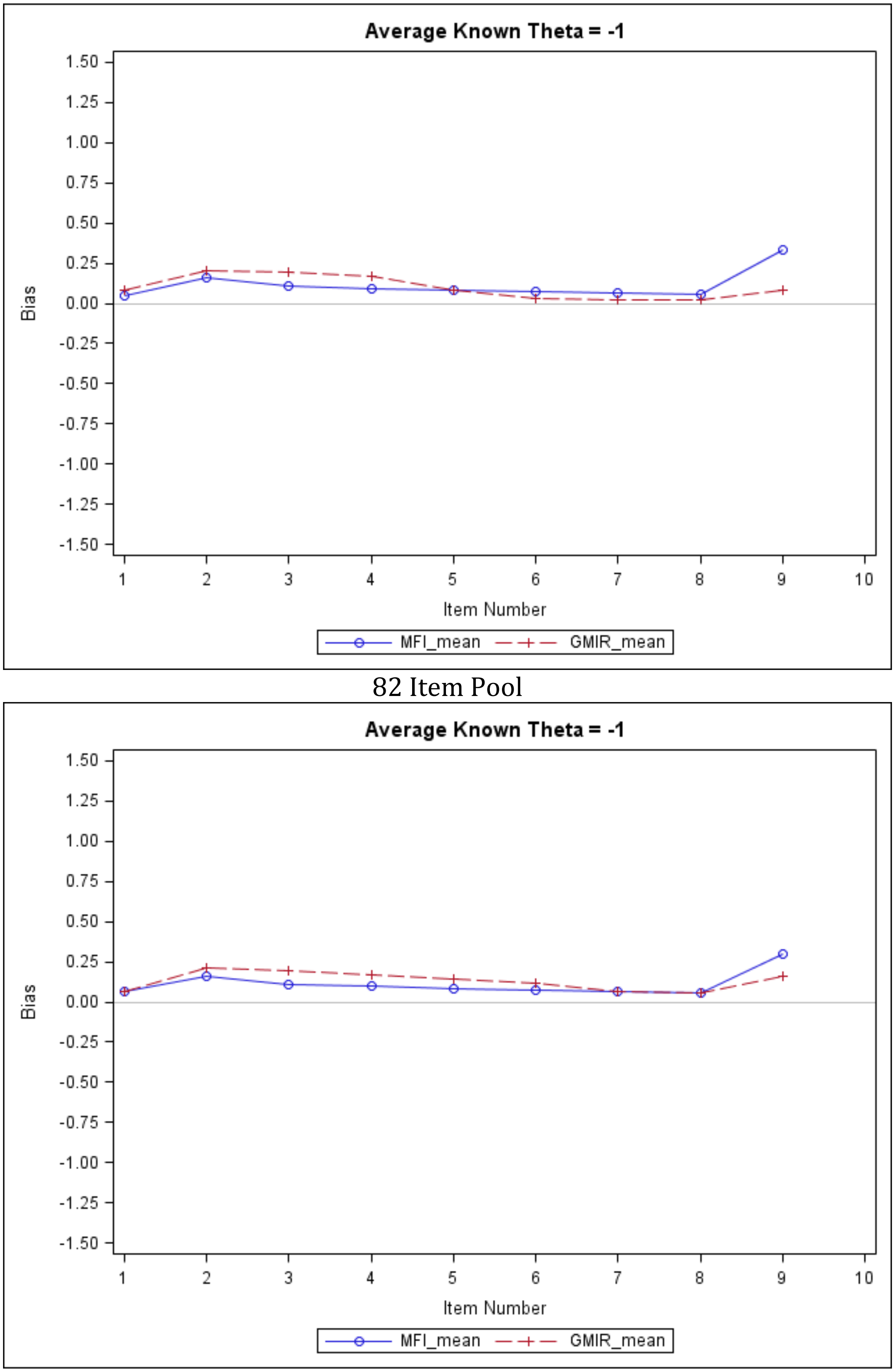
Figure 12A. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 5 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

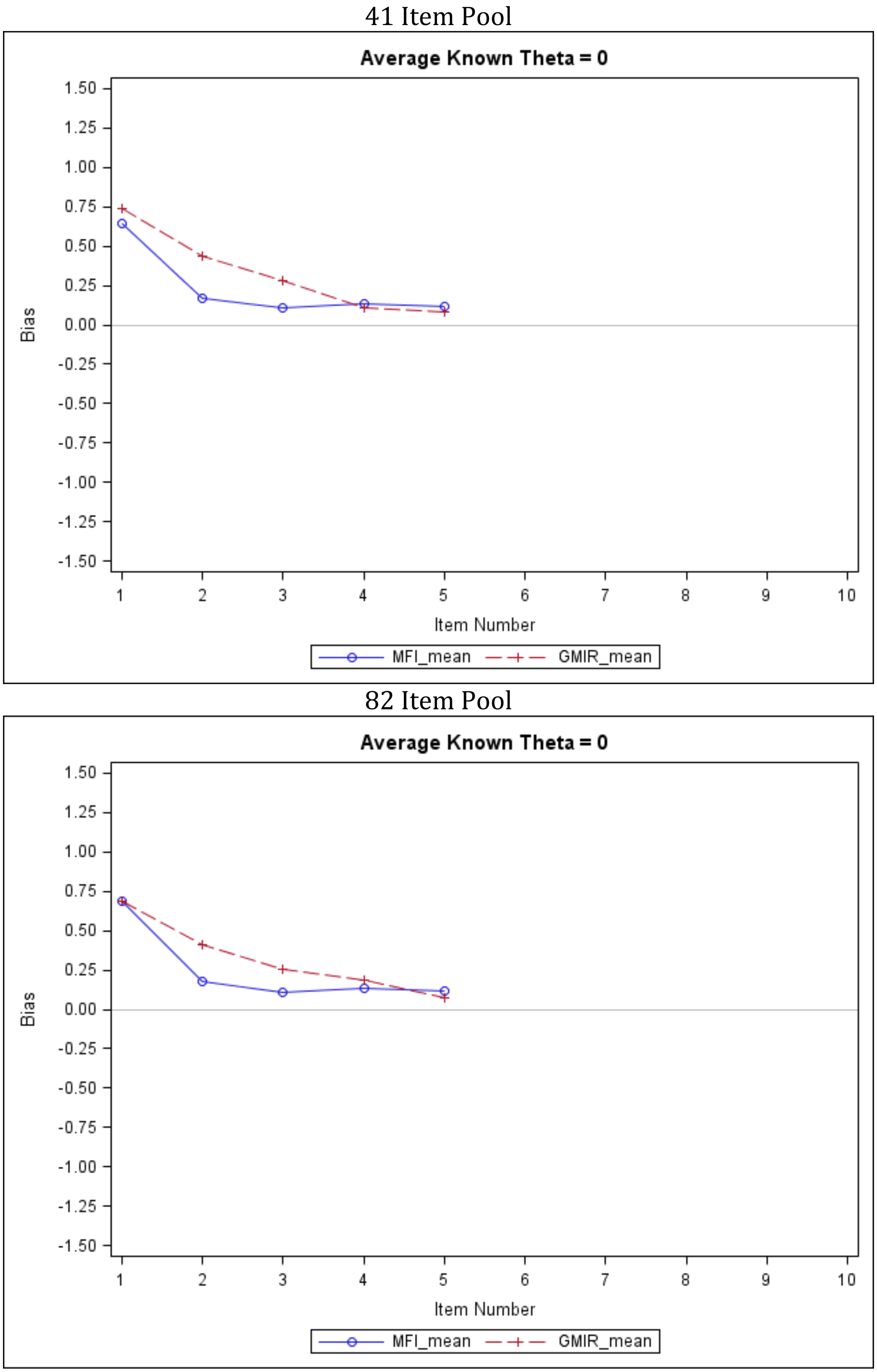
Figure 12B. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 7 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

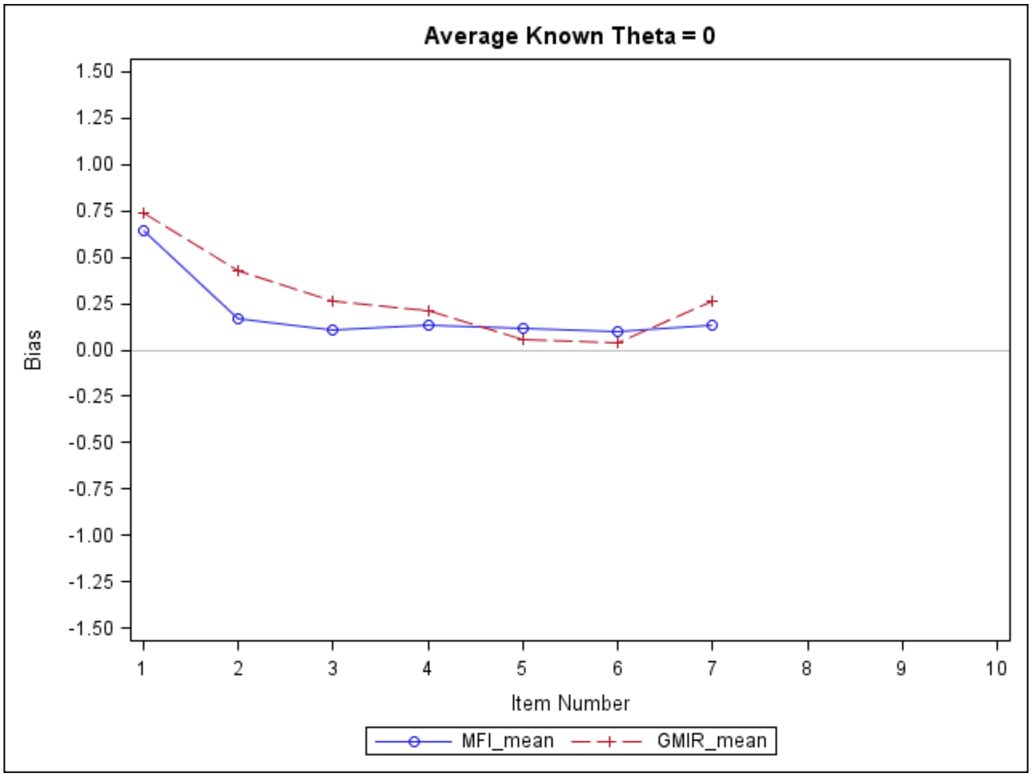
Figure 12C. Plots of Mean Bias Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 13A. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 5 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

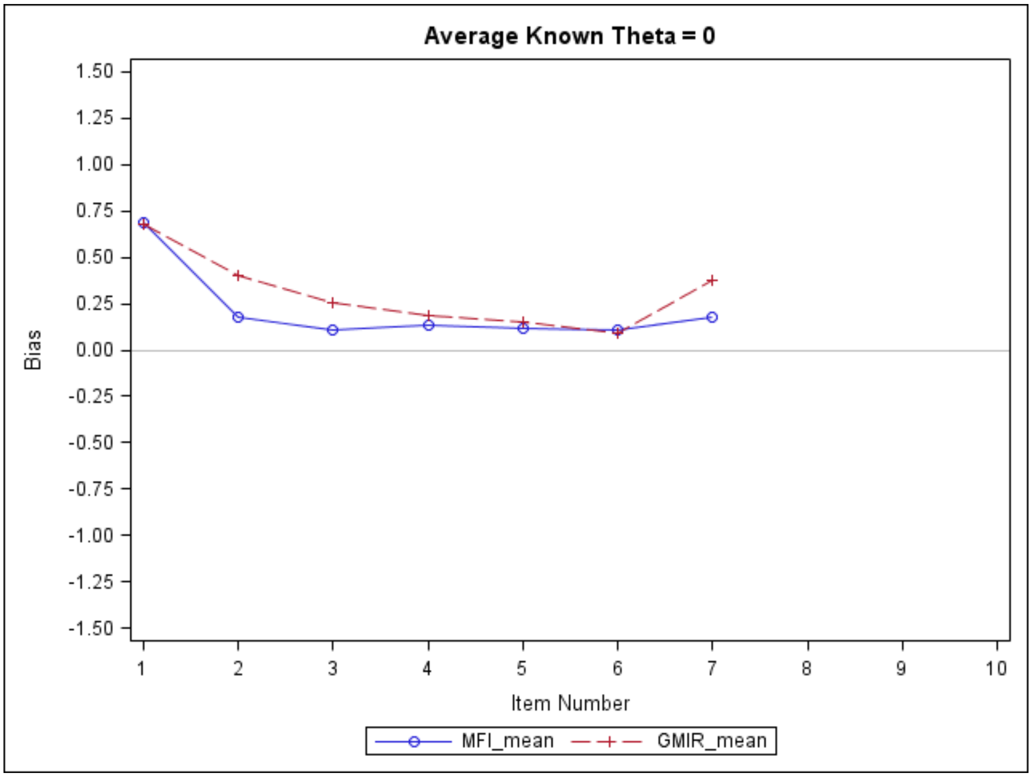
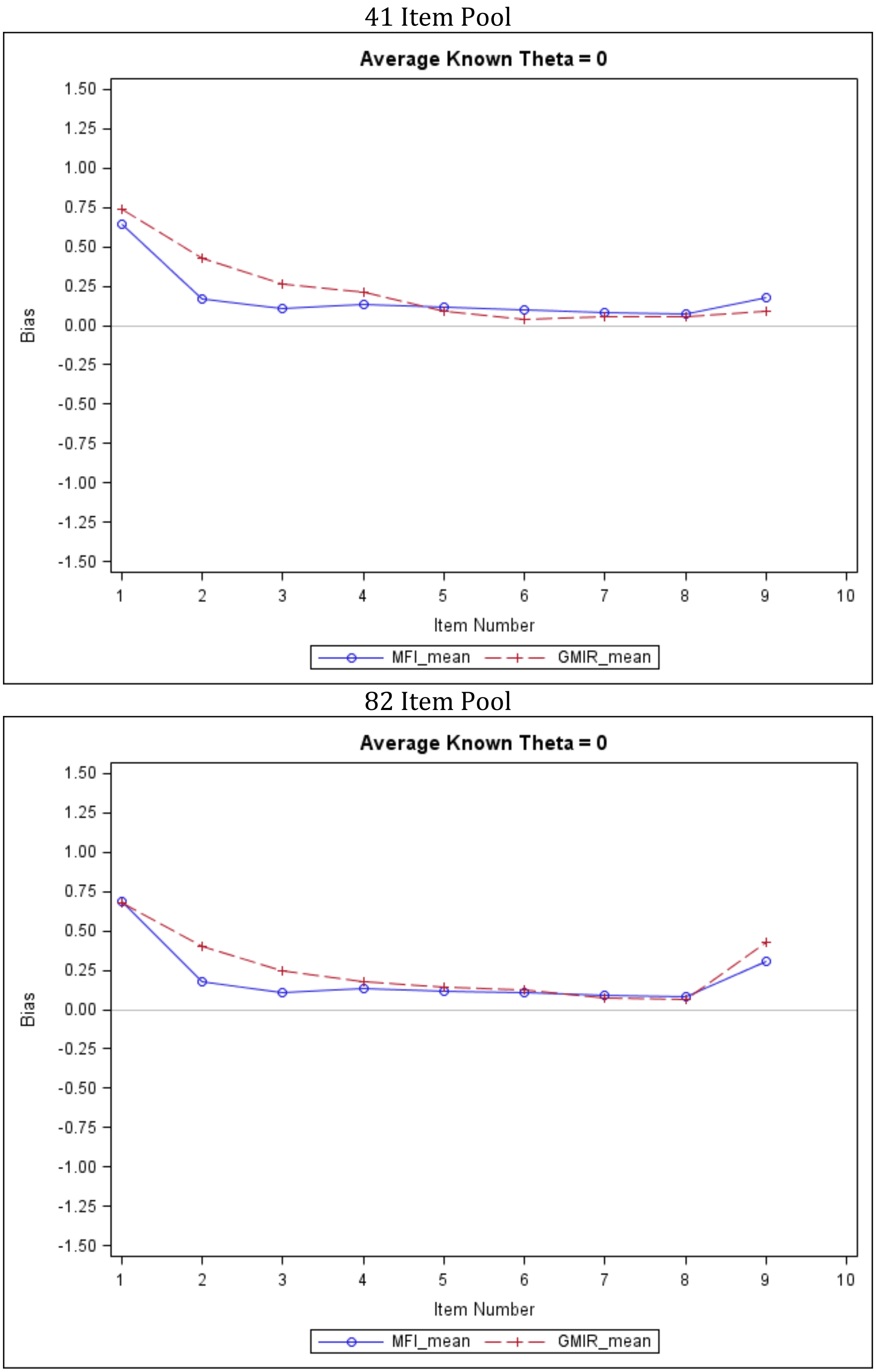


Figure 13B. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 7 Item Stopping Rule, and Normally Distributed Populations

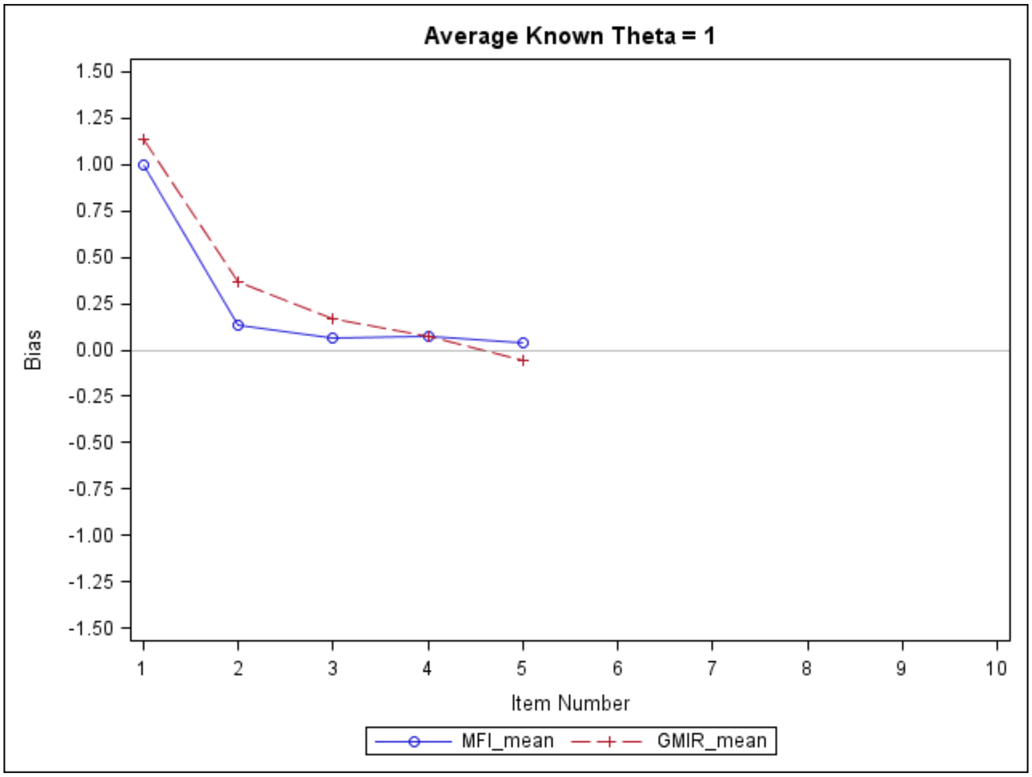


41 Item Pool

82 Item Pool

Figure 13C. Plots of Mean Bias Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

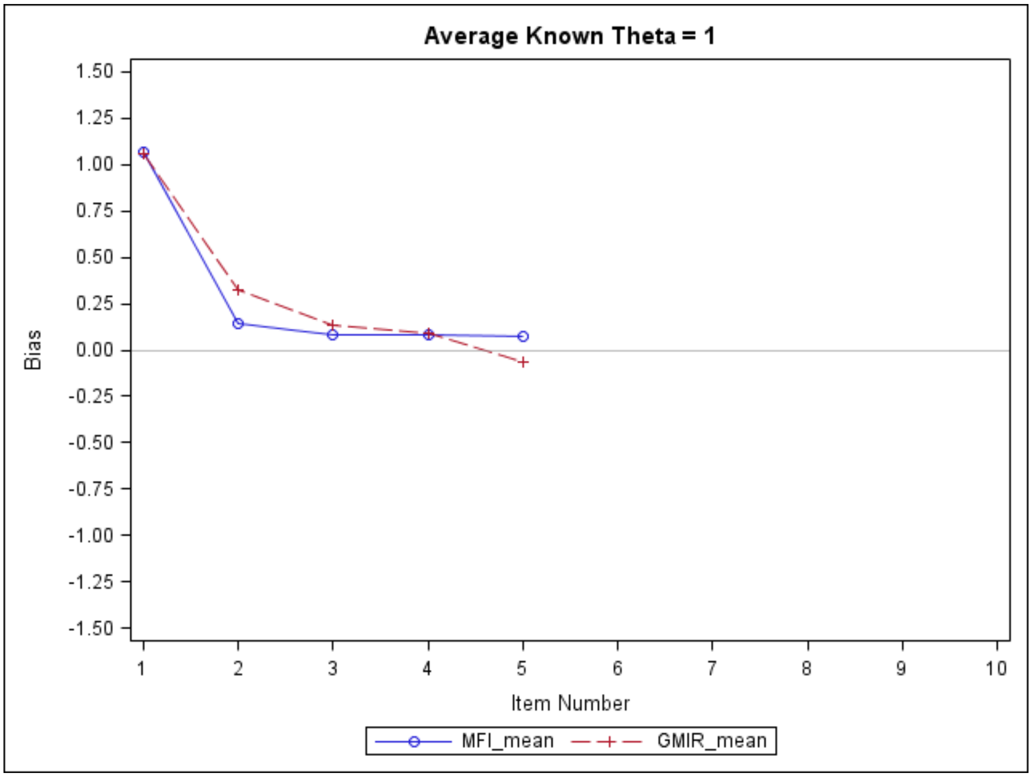
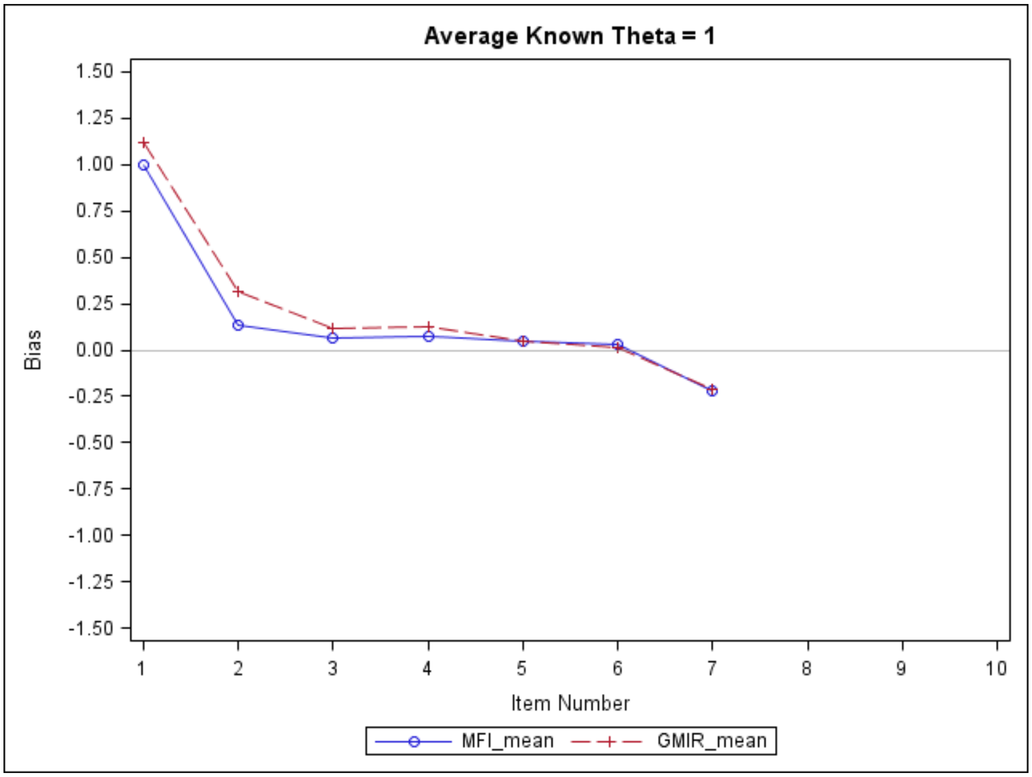


Figure 14A. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 5 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

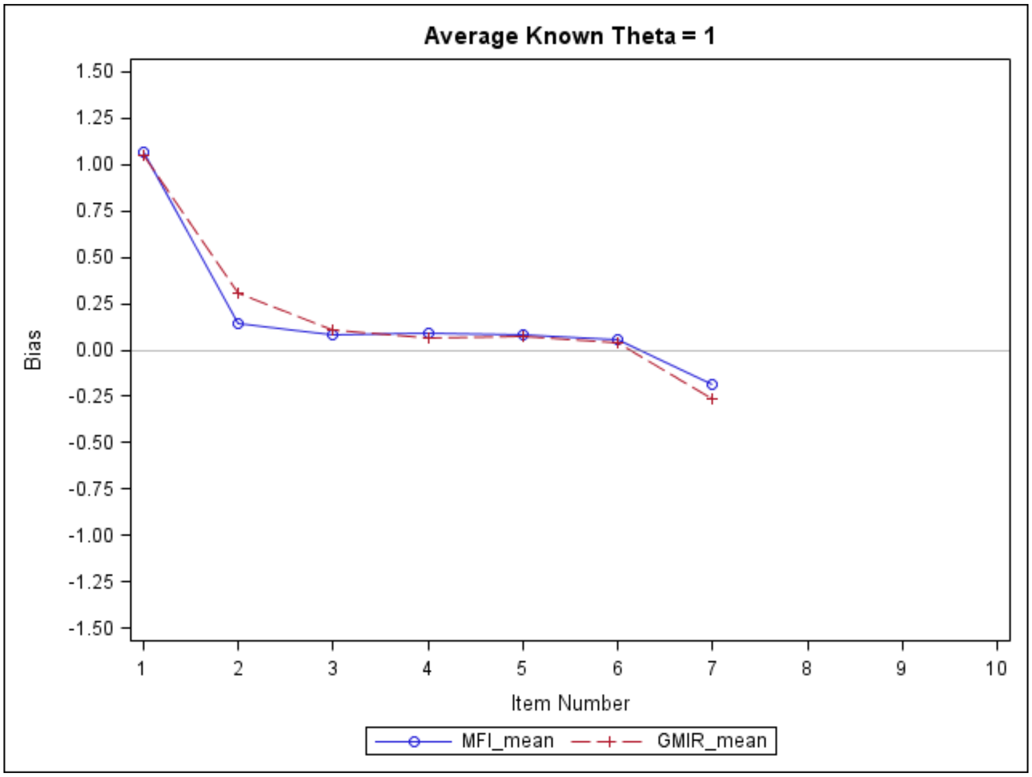
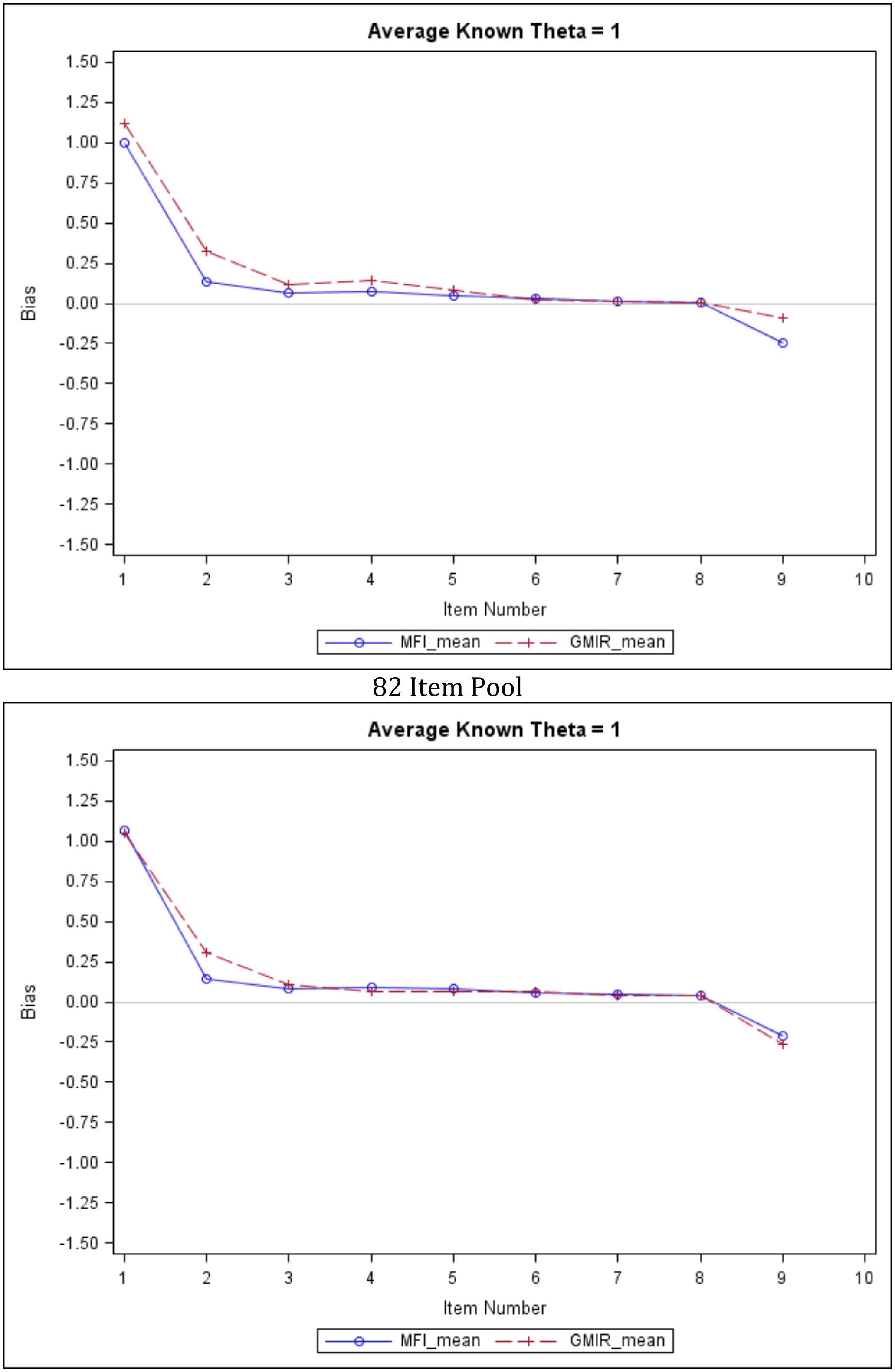


Figure 14B. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 7 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

Figure 14C. Plots of Mean Bias Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Populations

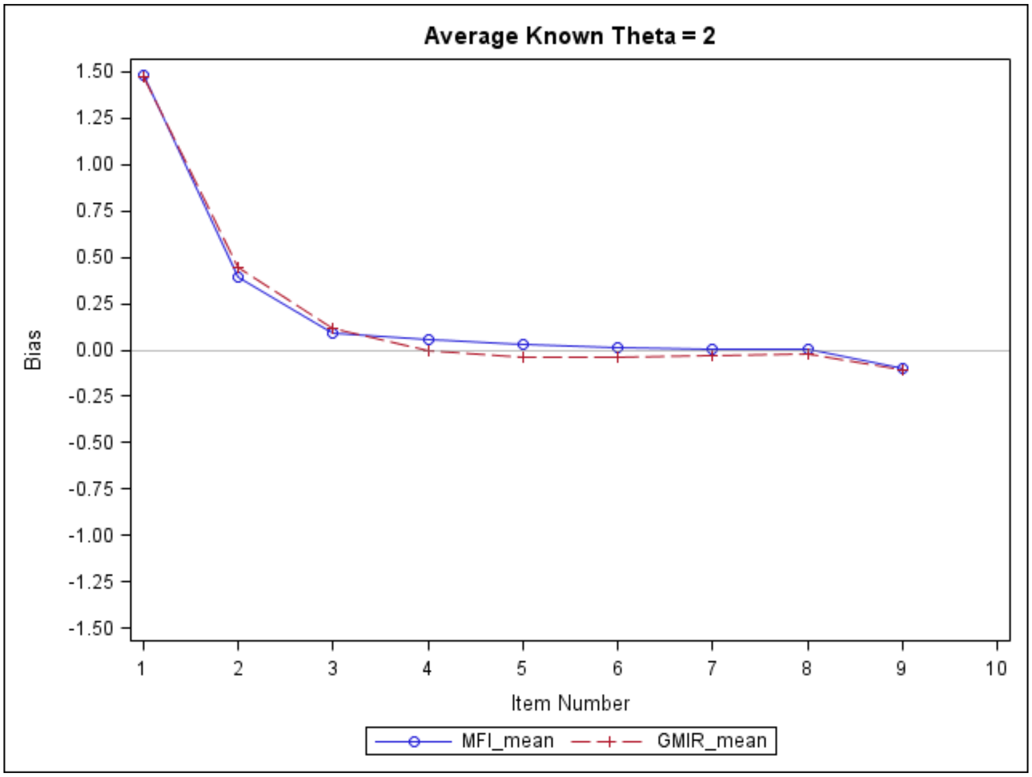
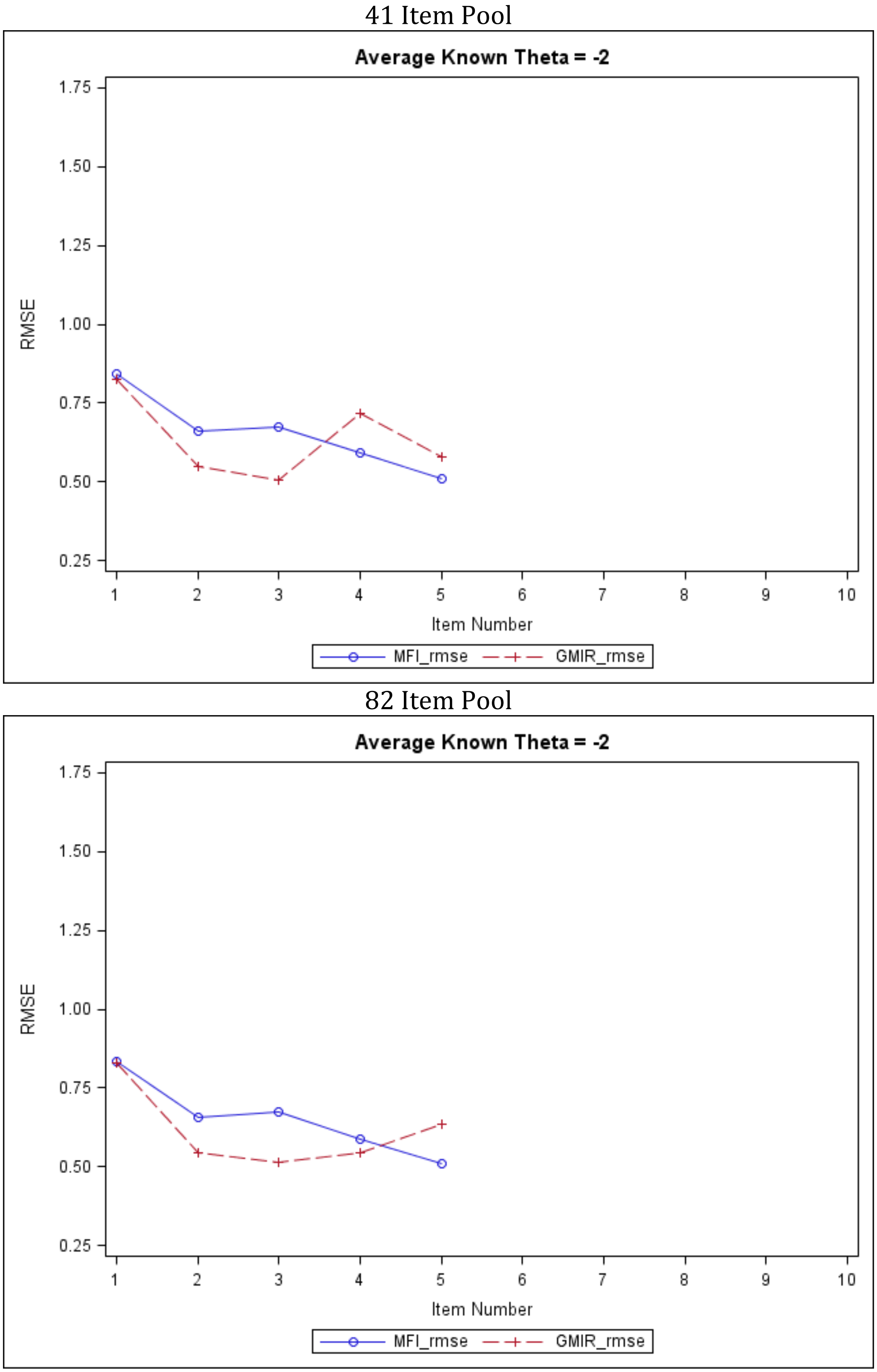


Figure 15. Plot of Mean Bias Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool

The conditional plots of RMSE by item for thetas -2, -1, 0, 1, and 2 are shown in Figures 16-21. The RMSE by item did not vary between the normally distributed population conditions and the negatively skewed population conditions except for simulees around theta=2. For theta groups -2, -1, 0, and 1 only the normal conditions are shown, the negatively skewed plots can be found in the Appendix A. Conditional plots of RMSE by item number for theta =-2 are shown in Figures 16. MFI and GMIR resulted in similar RMSE values at items 1 and 9, if given. At items 2-4, MFI RMSE values were higher, but then at items 5-8, GMIR RMSE values were higher. The differences were greater at items 5 the 41 item pool was used. For theta=-1, conditional plots of RMSE by item number are shown in Figures 17. RMSE values followed a similar pattern across the conditions depending on the test length.

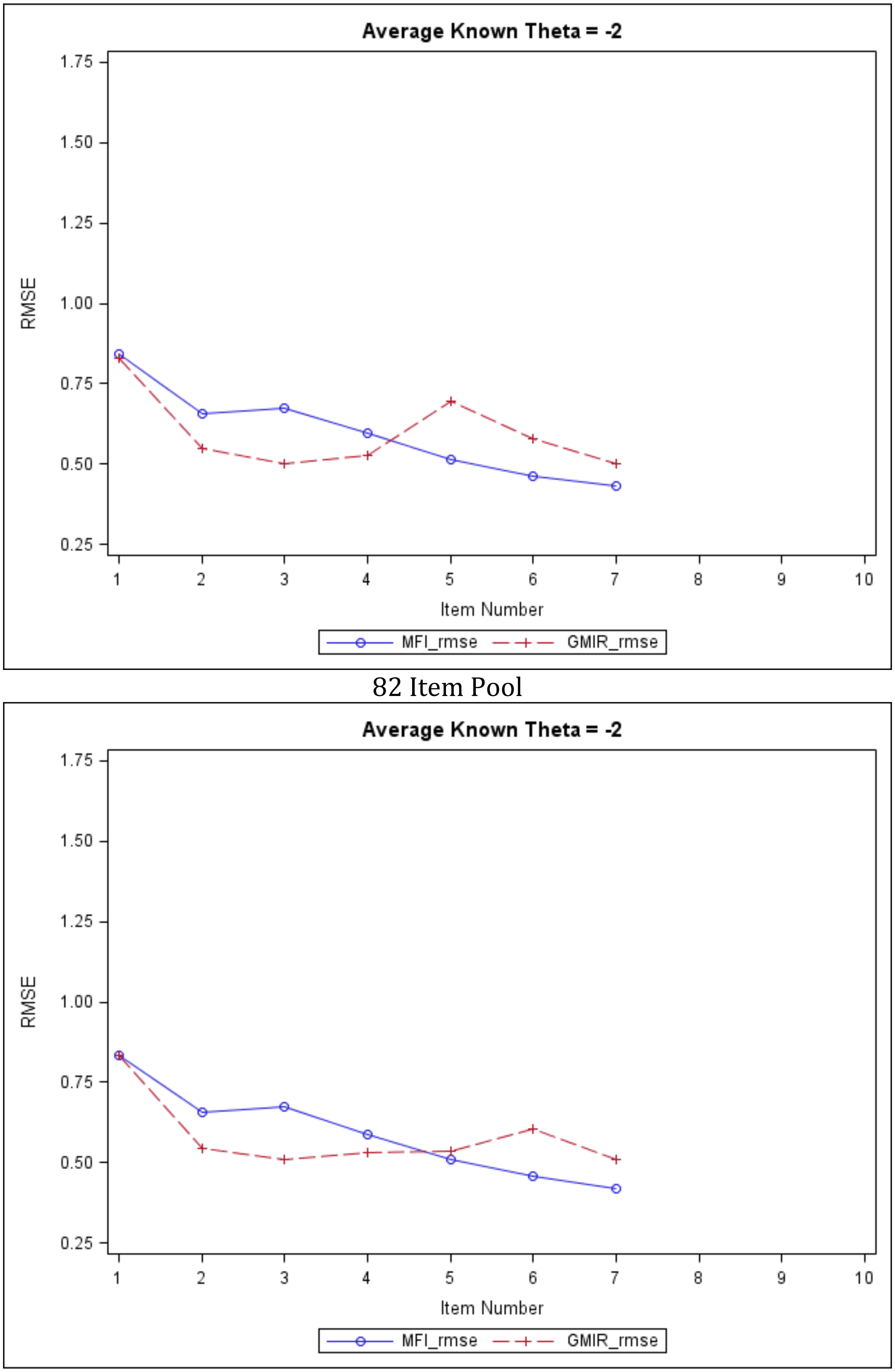
RMSE was slightly higher at items 1 and 2 when MFI was used. RMSE values were the same for the remainder of the items in the 5 item stopping rule conditions. In the 7 and 9 item stopping rule conditions, GMIR was slightly higher from items 3 to items 6 and 8 respectively. At the last item in the 7 and 9 item stopping rule conditions, MFI resulted in a higher RMSE than GMIR.

Figures 18 show the conditional plots of RMSE by item number for theta=0. Across conditions, GMIR results in a higher RMSE than MFI for items 2 and3. When the 5 item stopping rule is used, RMSE values are similar for item 5. When the 7 or 9 item stopping rules are used, MFI RMSE values are higher than GMIR values for the last 2-4 items. Conditions using the 41 item pool had a larger number of items at the end of the test where GMIR outperformed MFI than the 82 item pool conditions did. For theta=1, Figures 19 show the conditional plots of RMSE by item number. Conditions using the 41 item pool and GMIR resulted in higher RMSE values than MFI for the first few items: items 1-3 in the 5 item stopping rule condition, items, items 1-4 in the 7 item stopping rule condition, and items 1-5 in the 9 item stopping rule condition. Conditions using the 82 item pool did not have a difference in RMSE values between MFI and GMIR at item one, but GMIR conditions had higher values for the next few items: items 2-4, 2-5, and 2-5 for the 5, 7, and 9 item stopping rules respectively. In the 5 item stopping rule conditions, there was also a difference at item 5, but MFI resulted in a higher RMSE than GMIR. RMSE values at the end of the test for the 7 and 9 item conditions were comparable. Figures 20 and 21 show the RMSE values by item for theta=2. Using the 41 item pool and GMIR, RMSE values were slightly larger for items 1 and 2. The differences were larger for the negatively skewed population conditions than the normal population. The largest differences were found using the 5 item stopping rule. The 82 item conditions varied by population distribution, using the negatively skewed population distribution and GMIR resulted in similar first items and then slightly higher RMSE values for all the middle items. The normally distributed population and 82 item conditions resulted in similar values using GMIR and MFI.


41 Item Pool

82 Item Pool

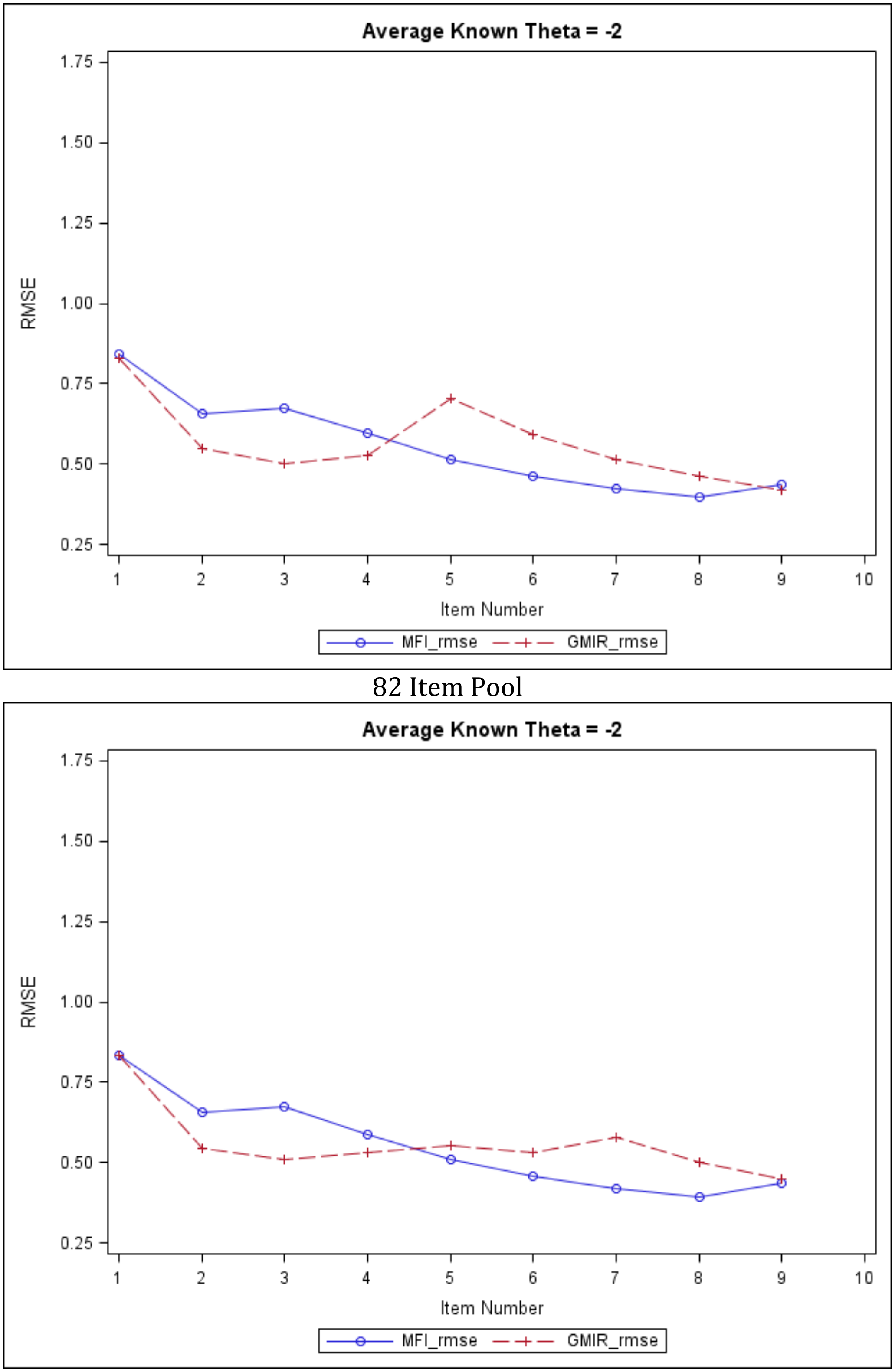
Figure 16A. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 16B. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Populations

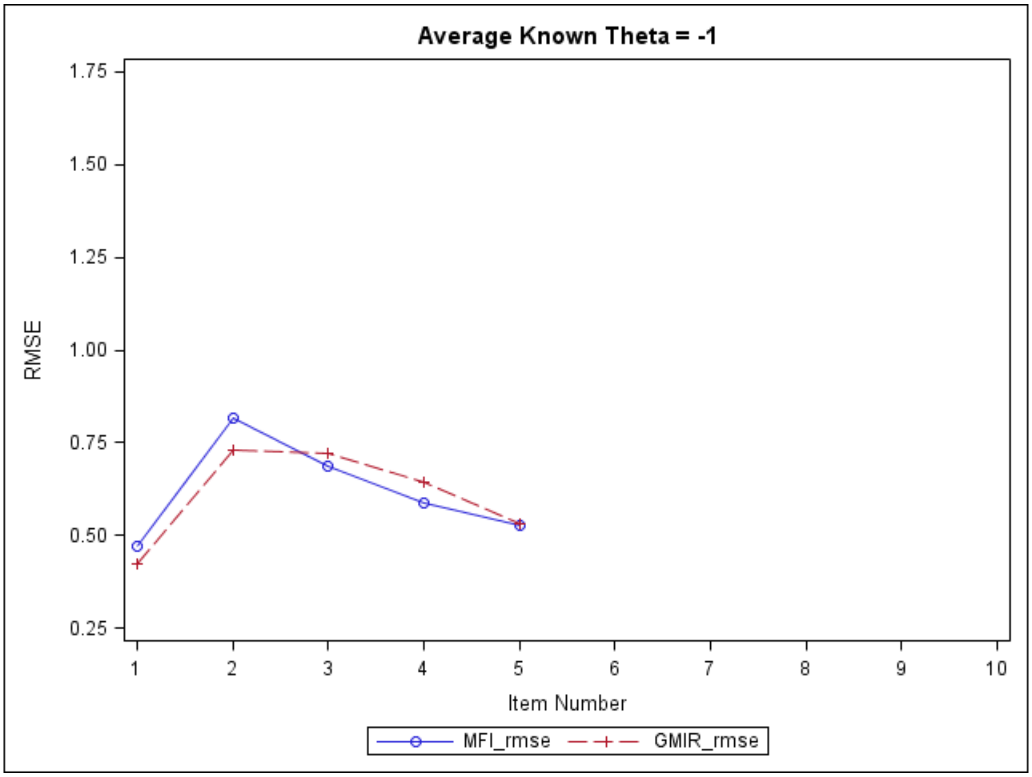
41 Item Pool



82 Item Pool

Figure 16C. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

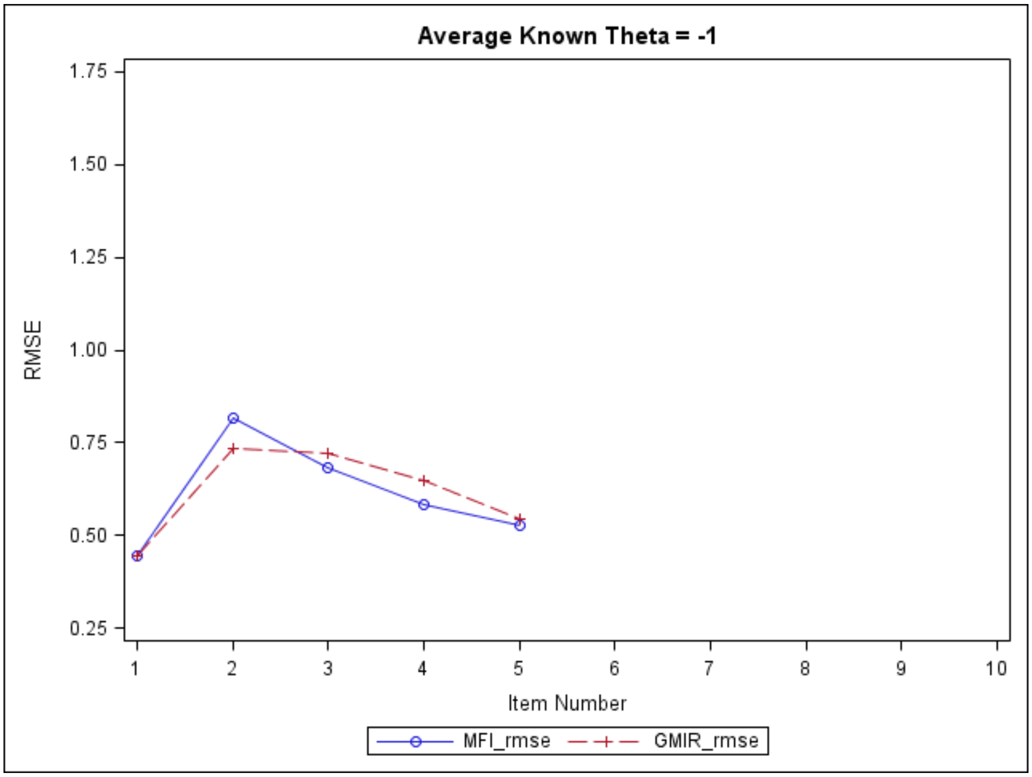
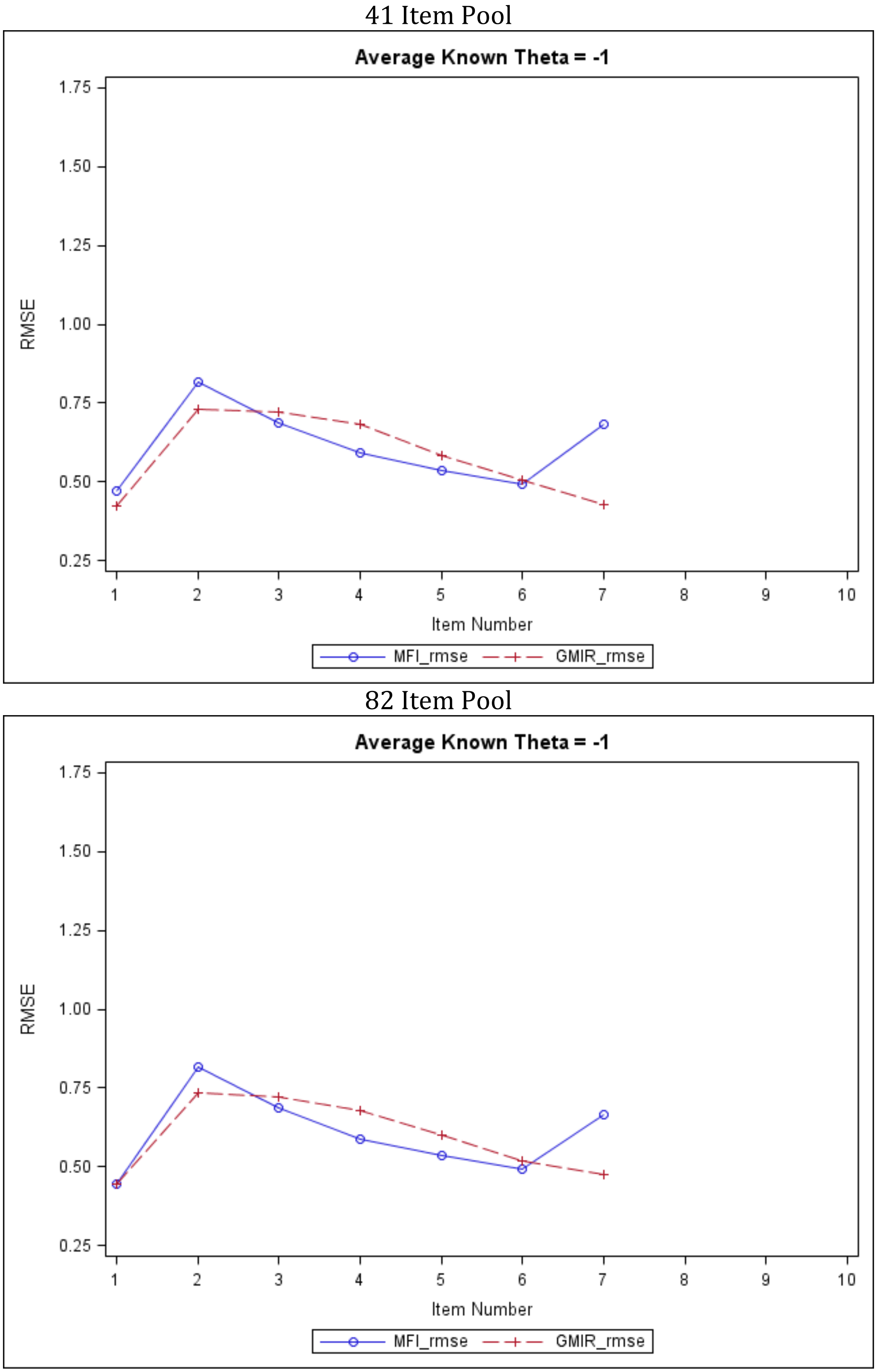


Figure 17A. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1, the 5 Item Stopping Rule, and Normally Distributed Populations

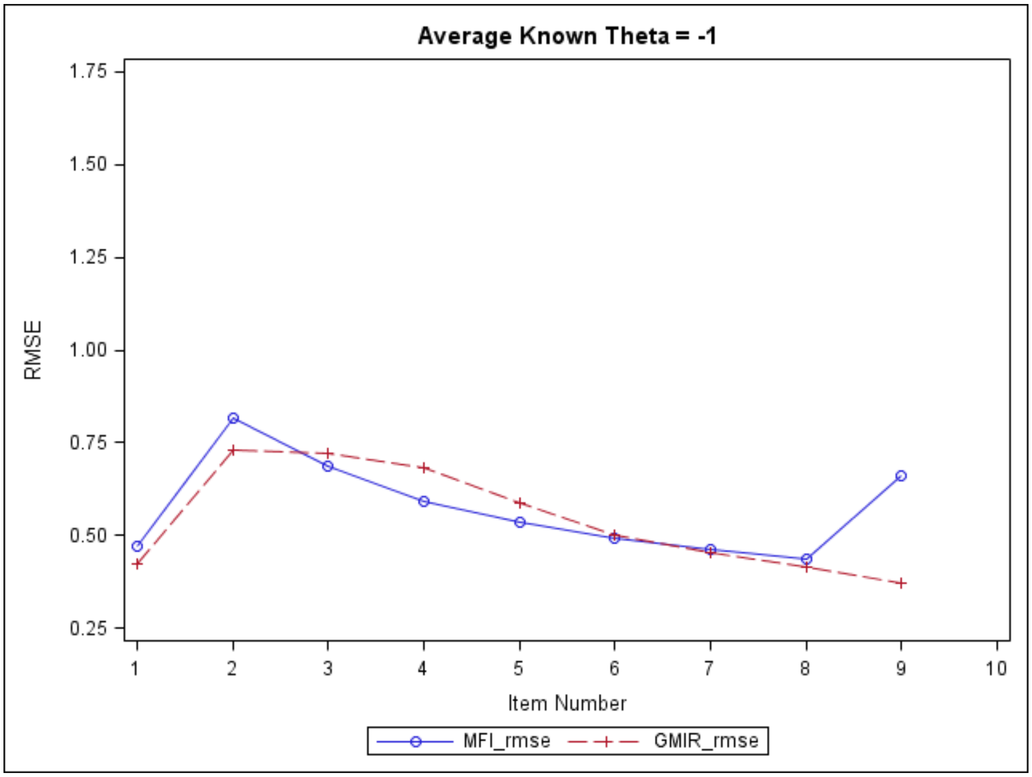


41 Item Pool

82 Item Pool

Figure 17B. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1, the 7 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

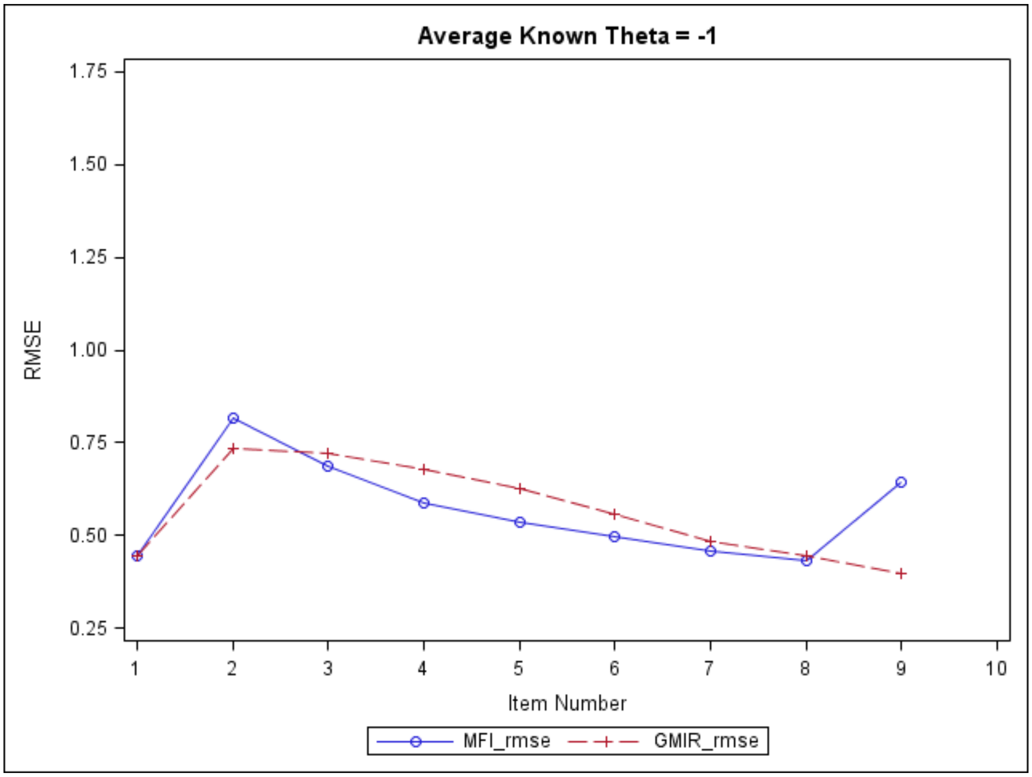
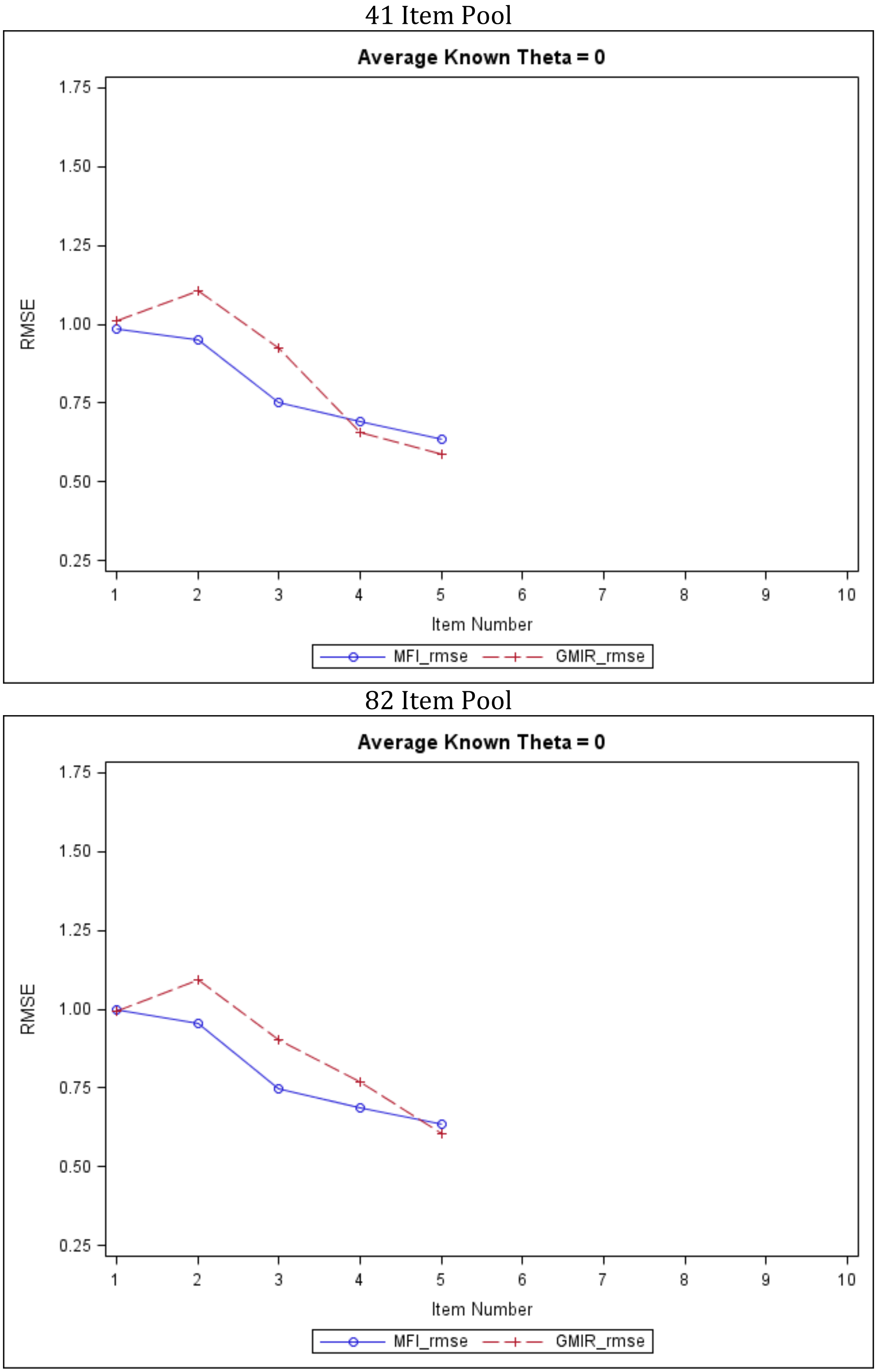


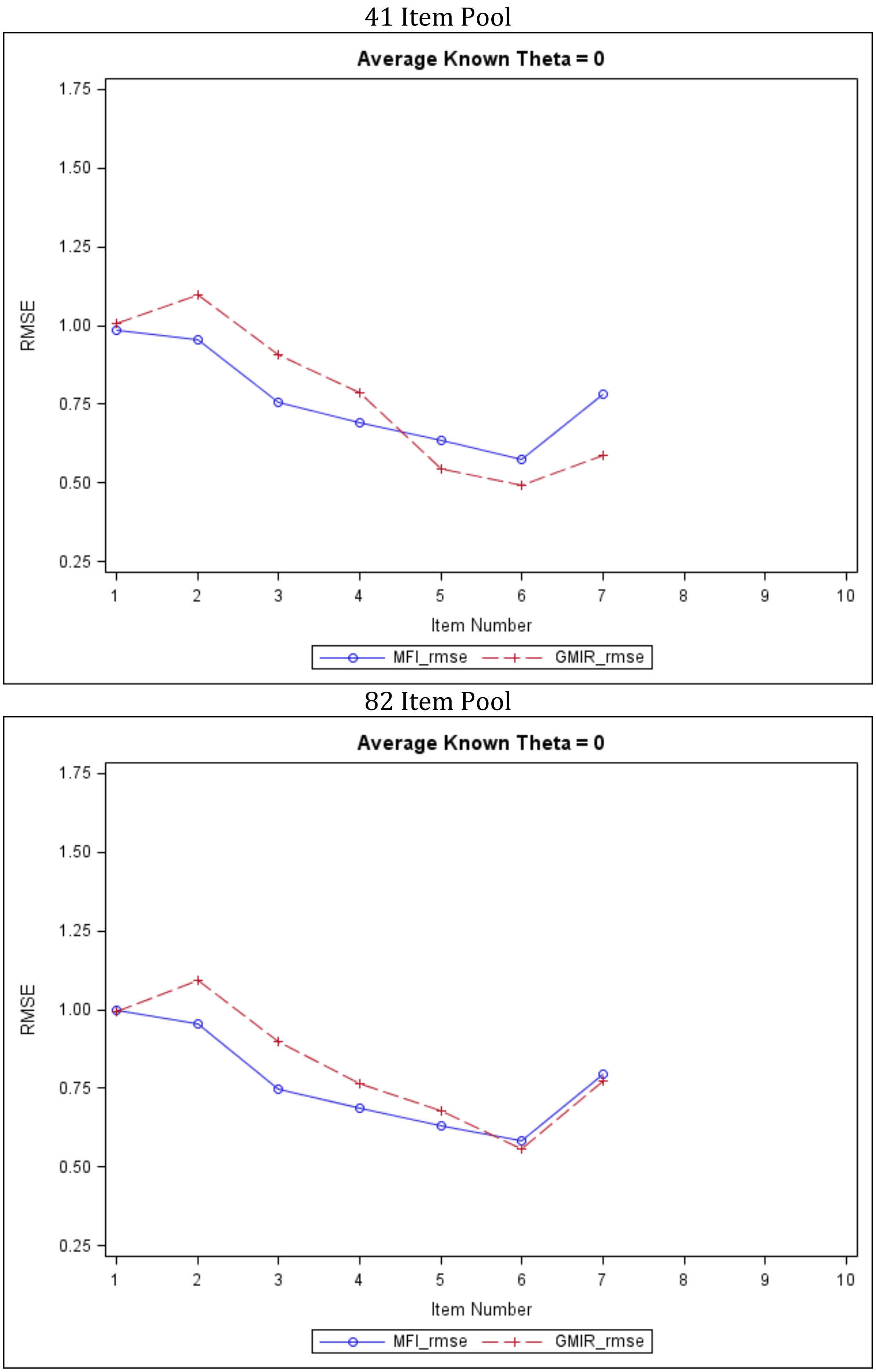
Figure 17C. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

82 Item Pool

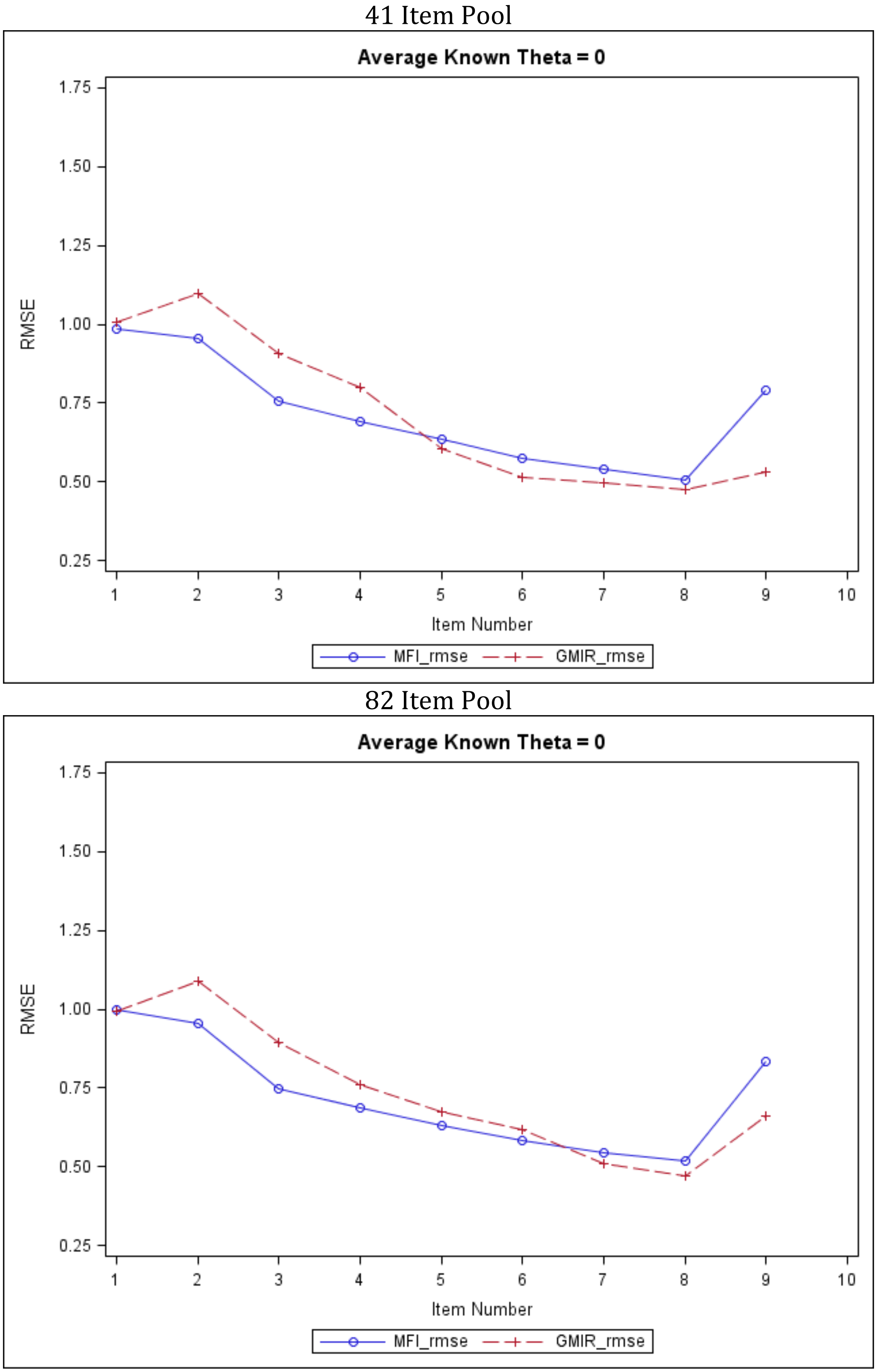
Figure 18A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 5 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

82 Item Pool

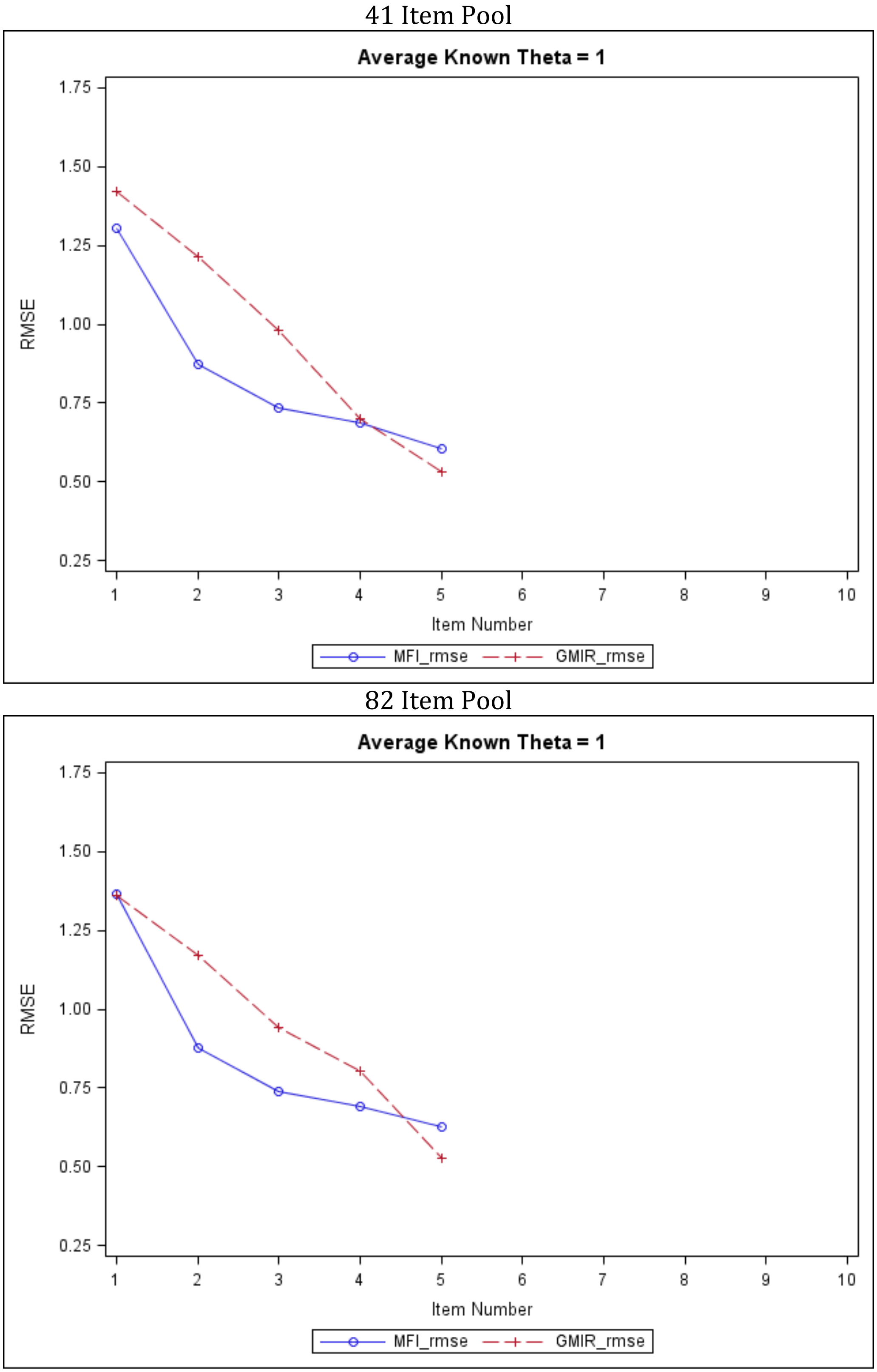
Figure 18B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 7 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

82 Item Pool

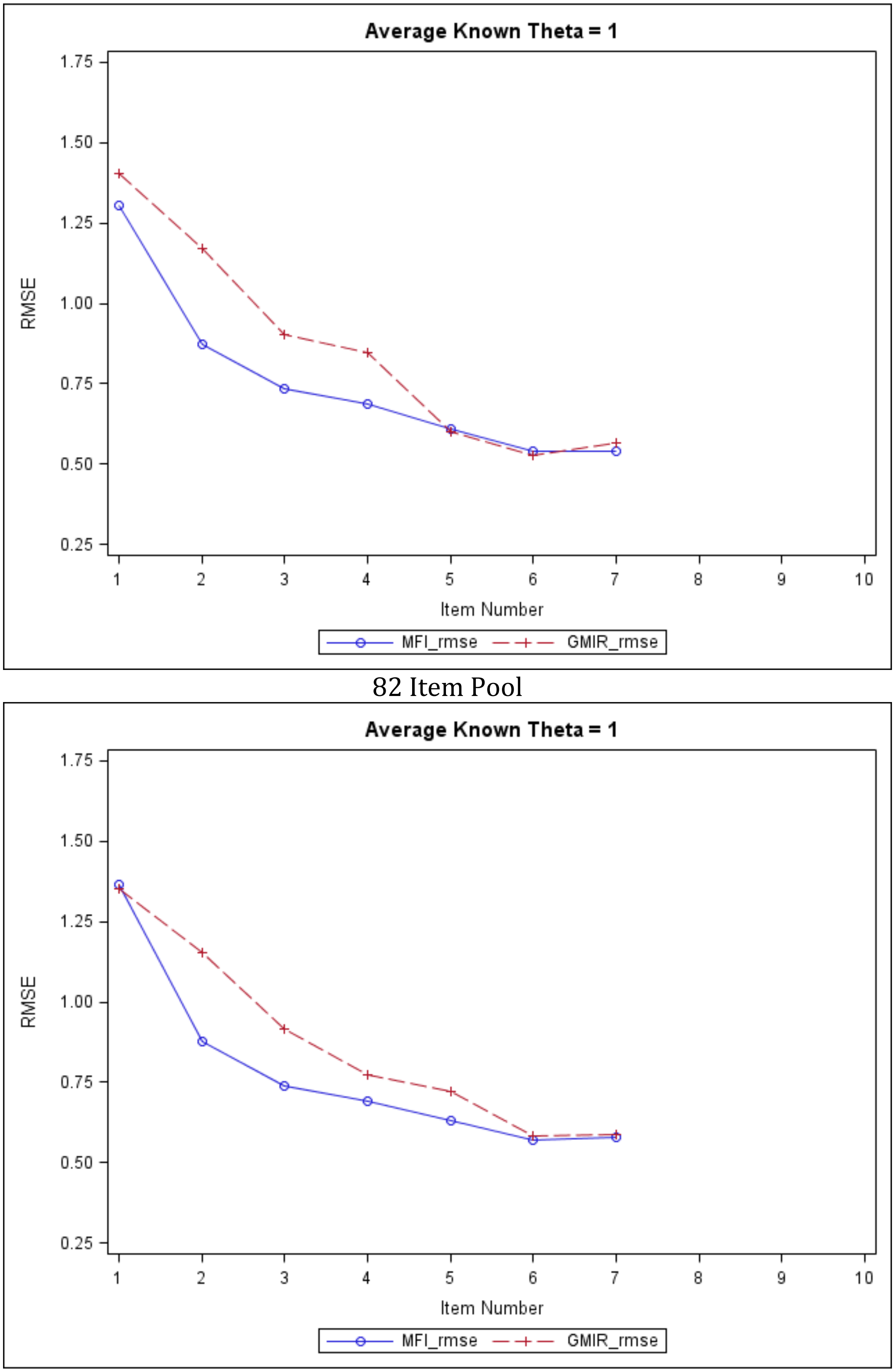
Figure 18C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

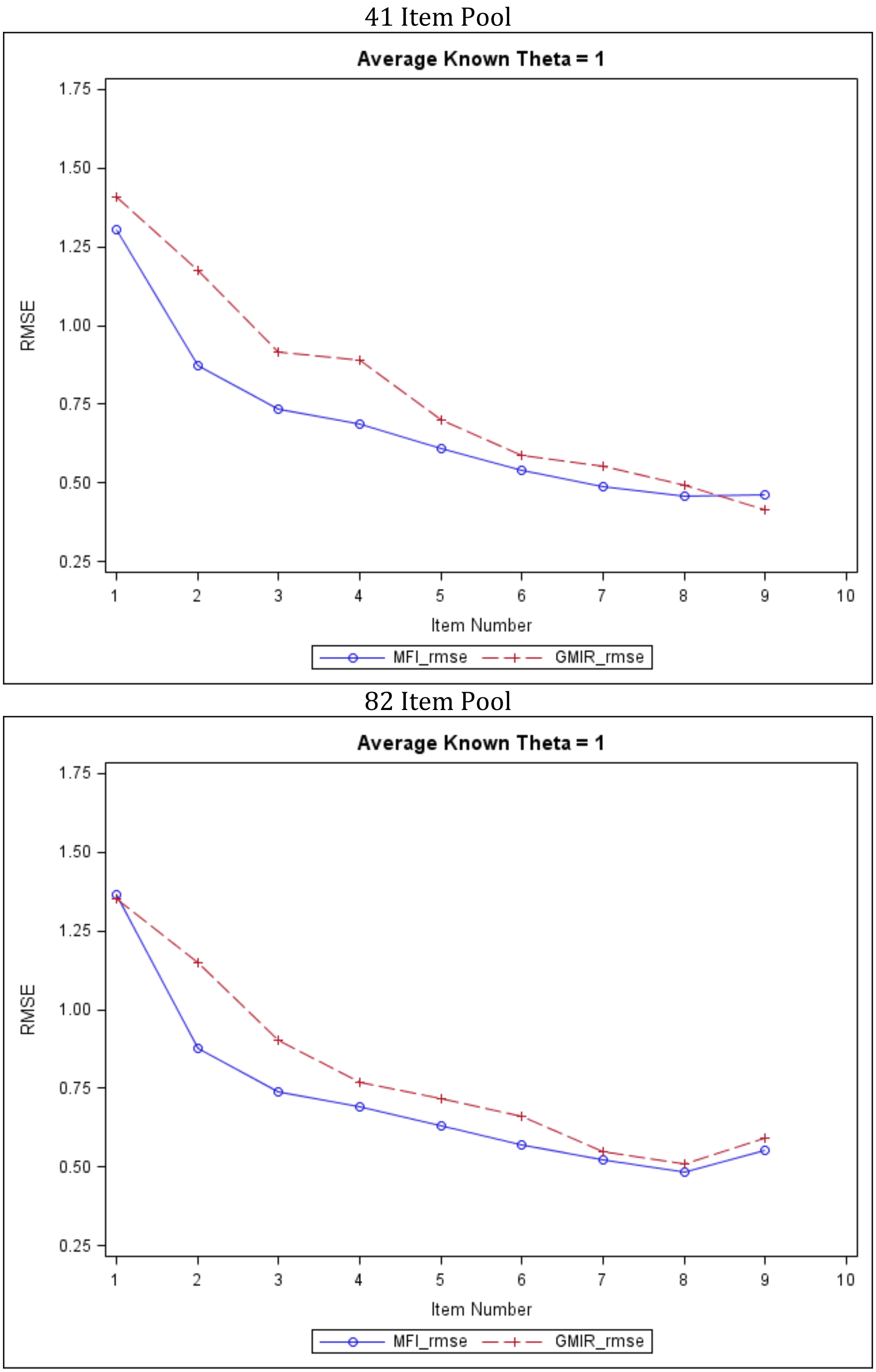
Figure 19A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 5 Item Stopping Rule, and Normally Distributed Populations

41 Item Pool



82 Item Pool

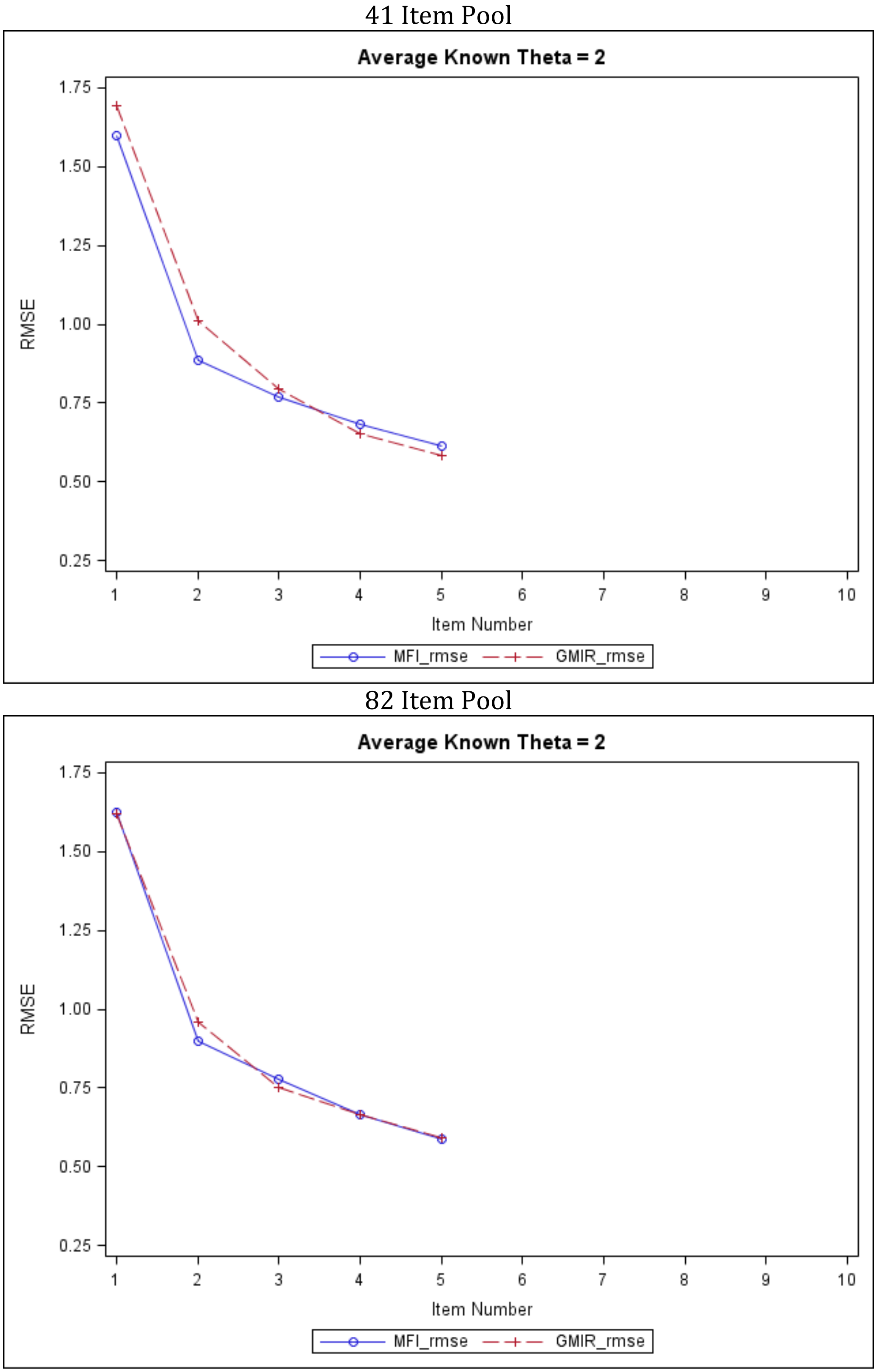
Figure 19B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 7 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

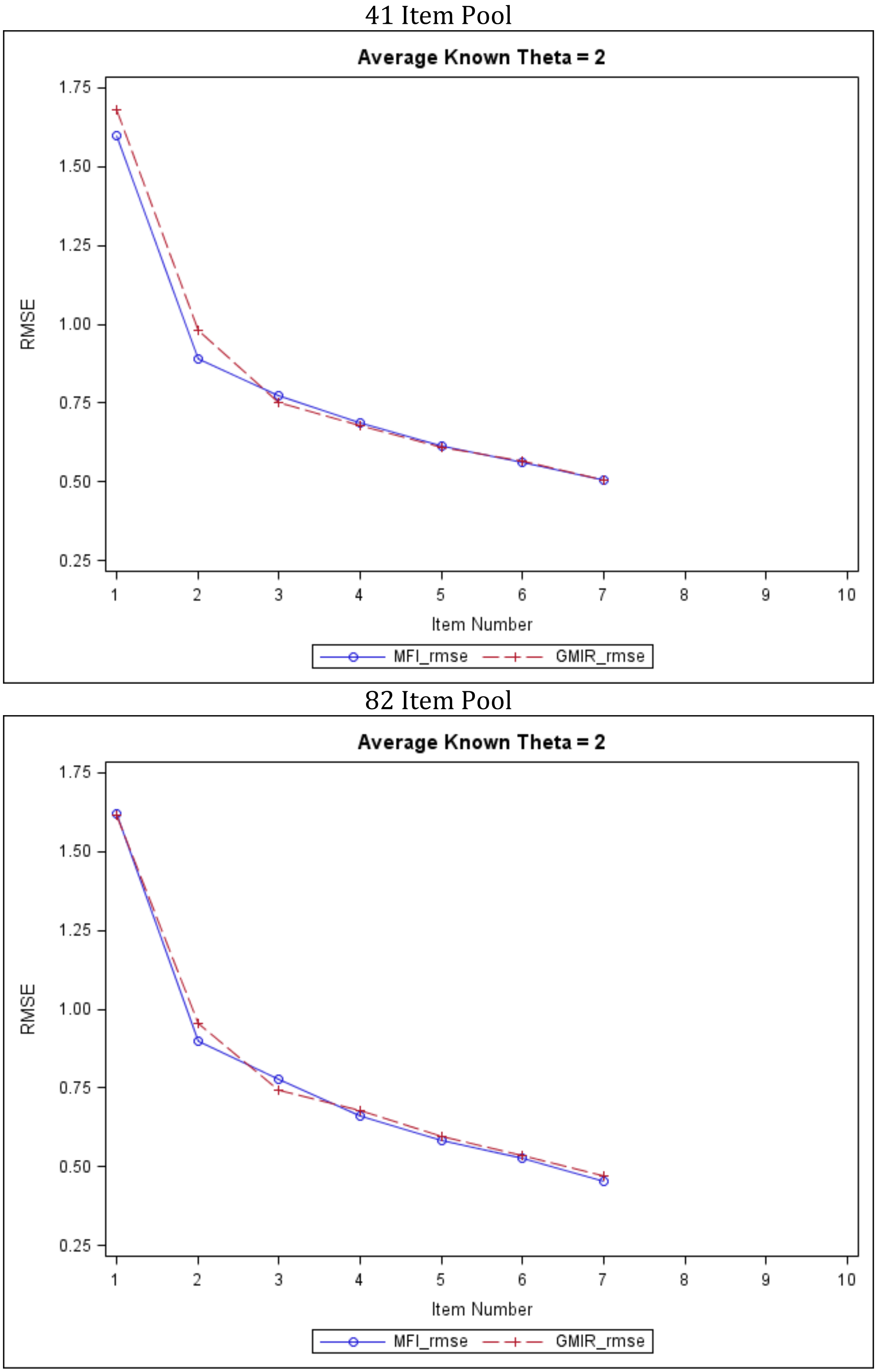
82 Item Pool

Figure 19C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

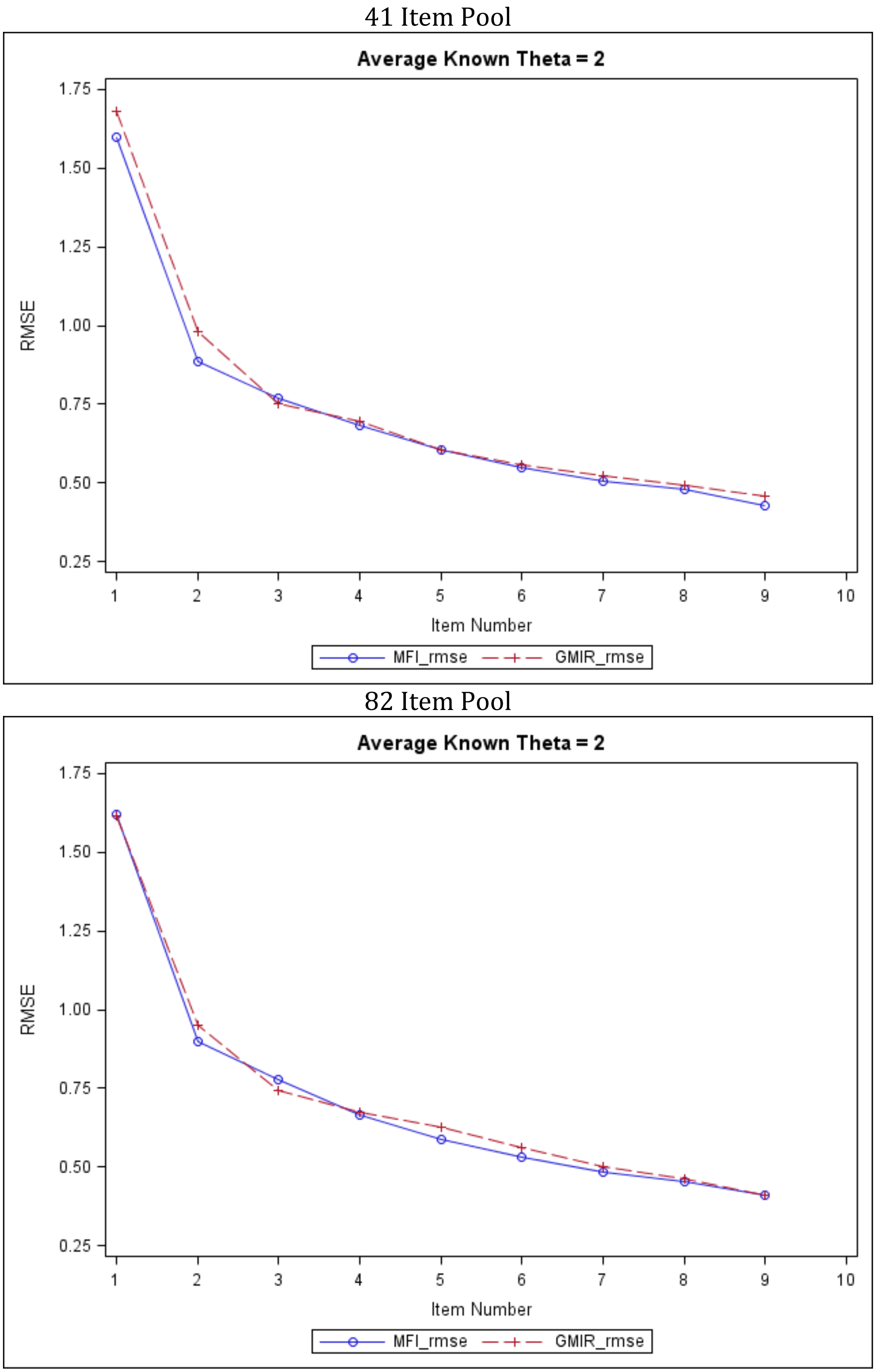
Figure 20A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 5 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

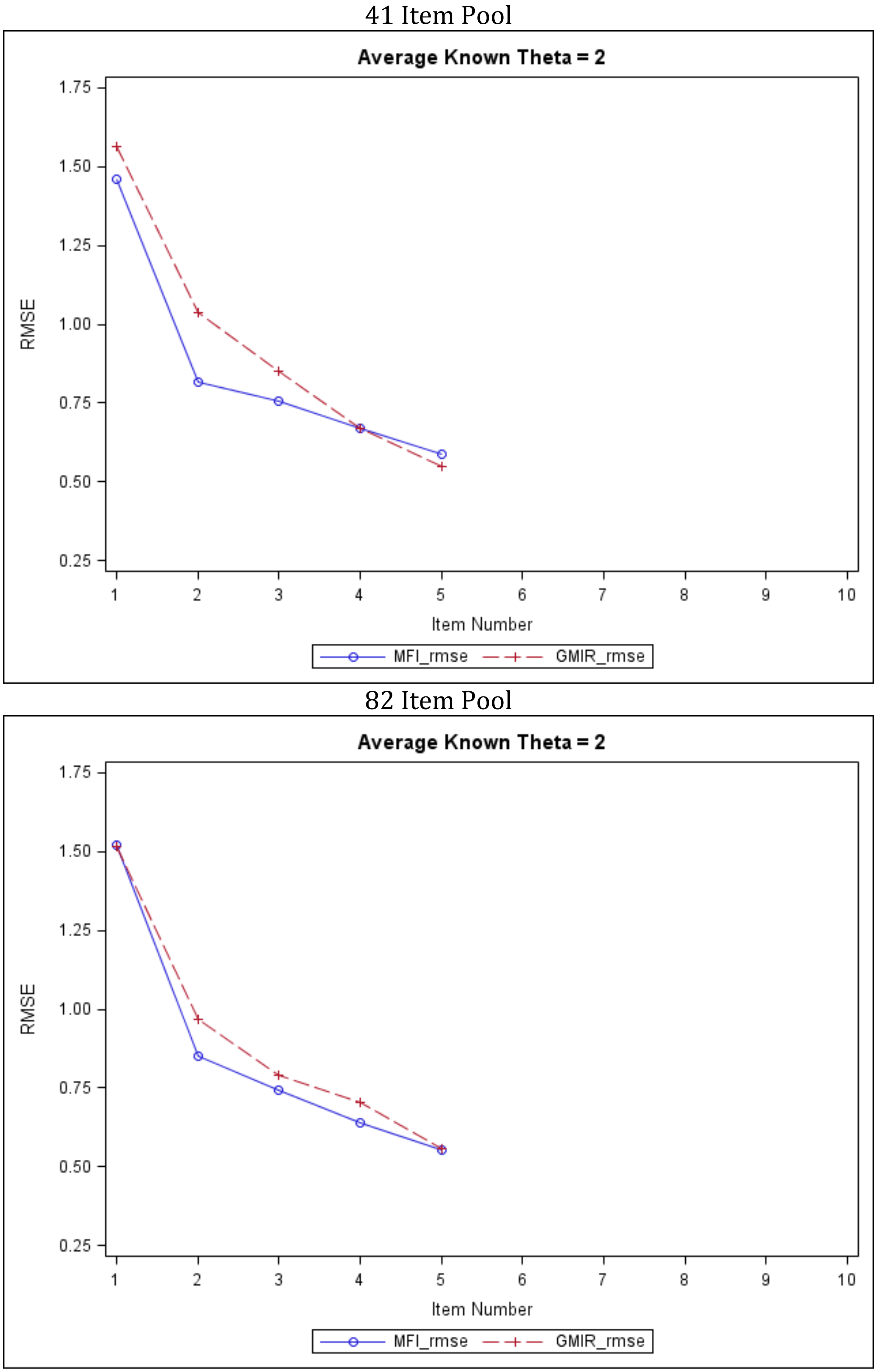
82 Item Pool

Figure 20B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 7 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 20C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, and Normally Distributed Populations

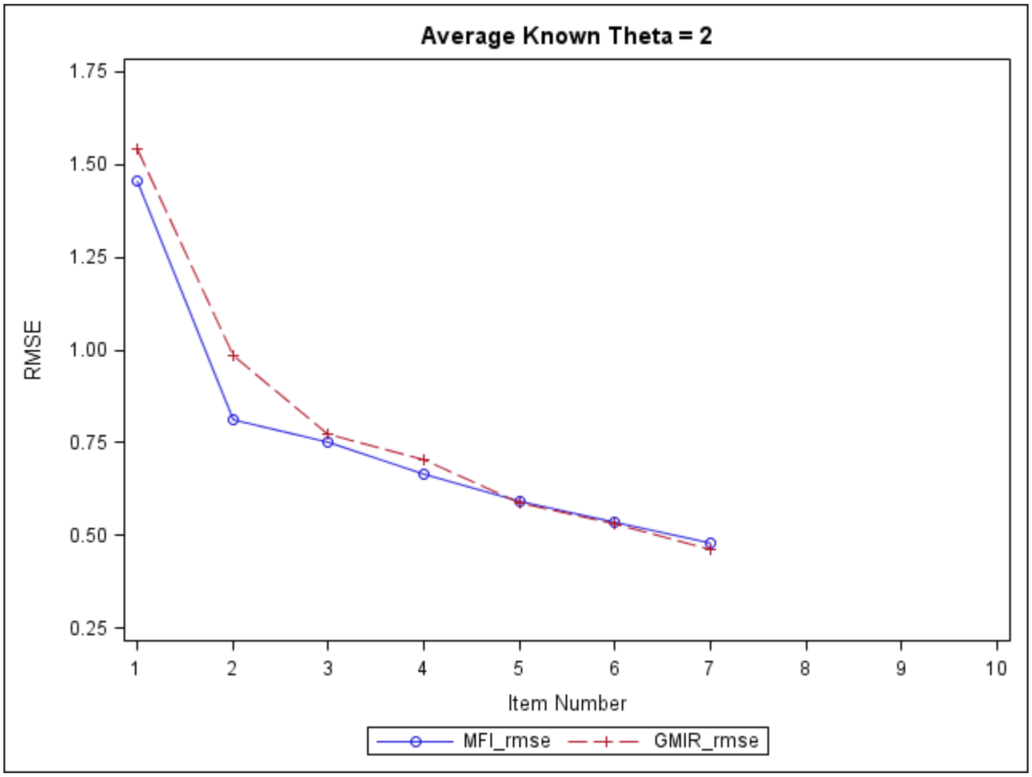


41 Item Pool

82 Item Pool

Figure 21A. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 5 Item Stopping Rule, and Negatively Skewed Distributed Populations

41 Item Pool



82 Item Pool

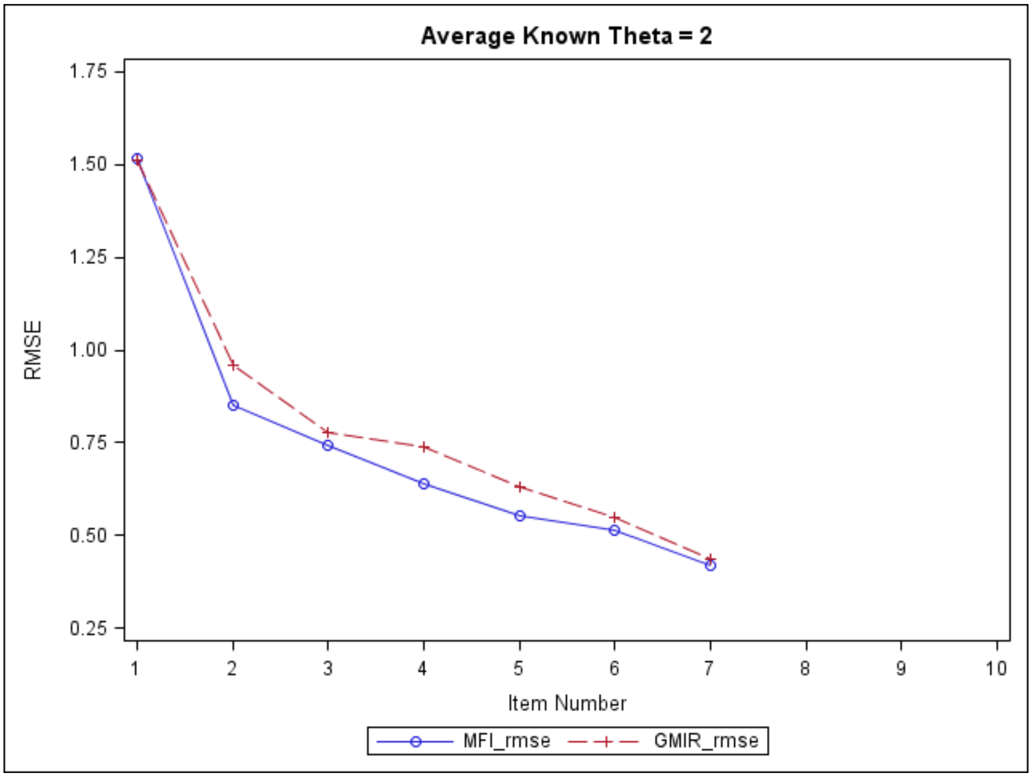
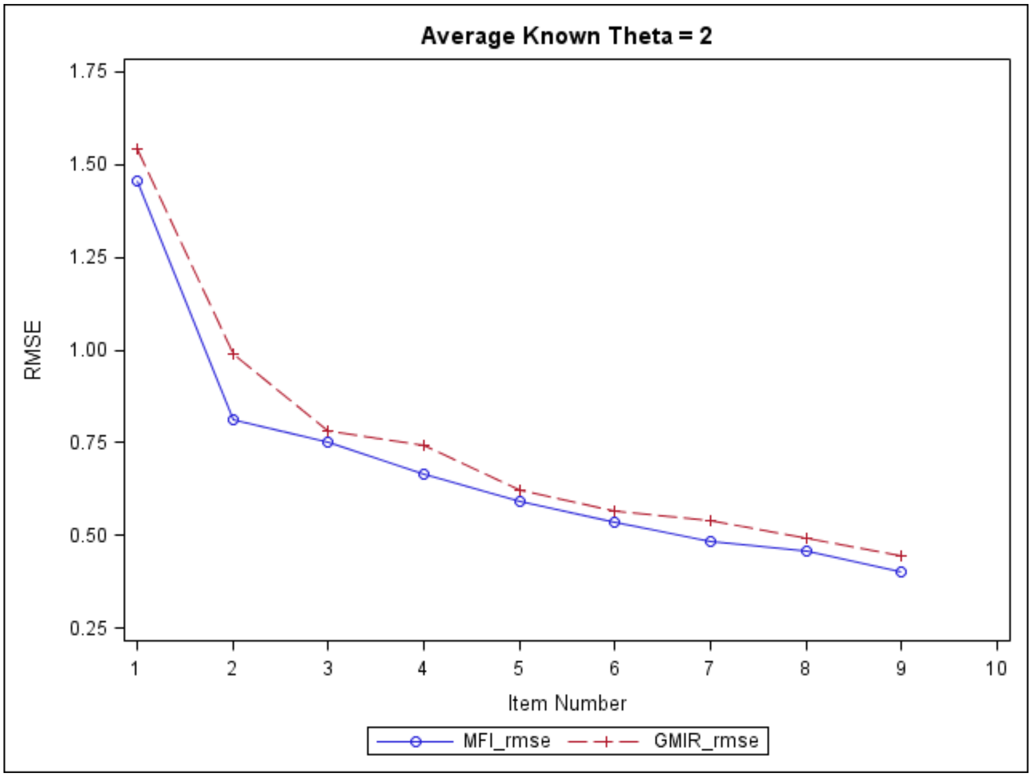


Figure 21B. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 7 Item Stopping Rule, and Negatively Skewed Distributed Populations

41 Item Pool



82 Item Pool

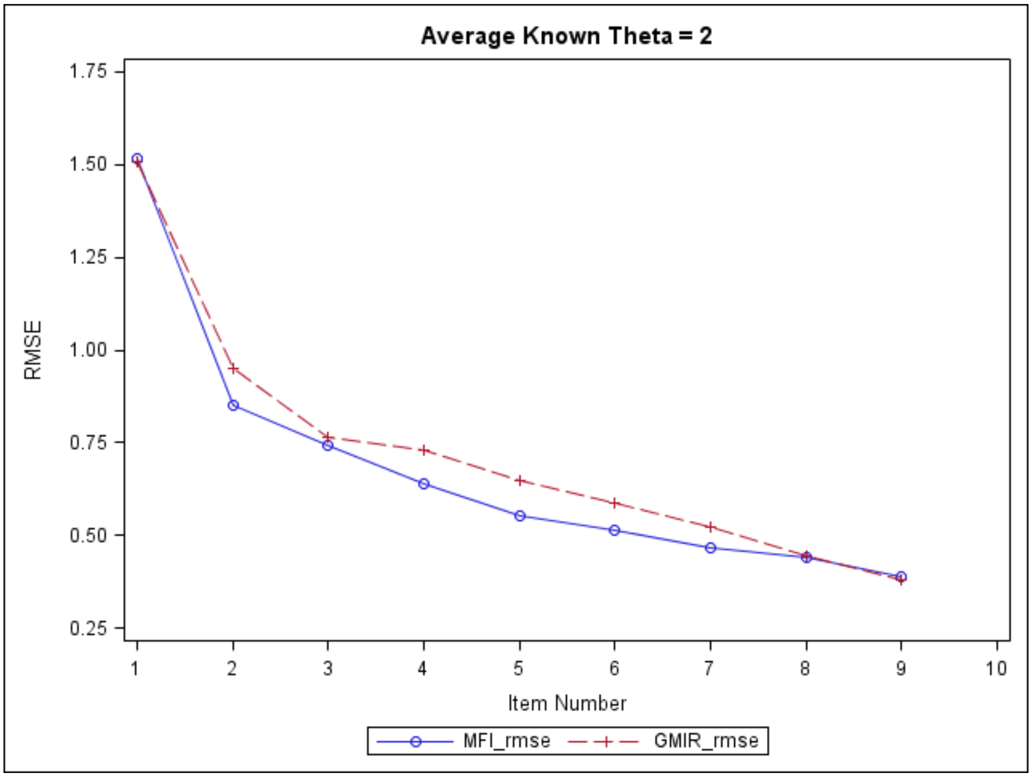
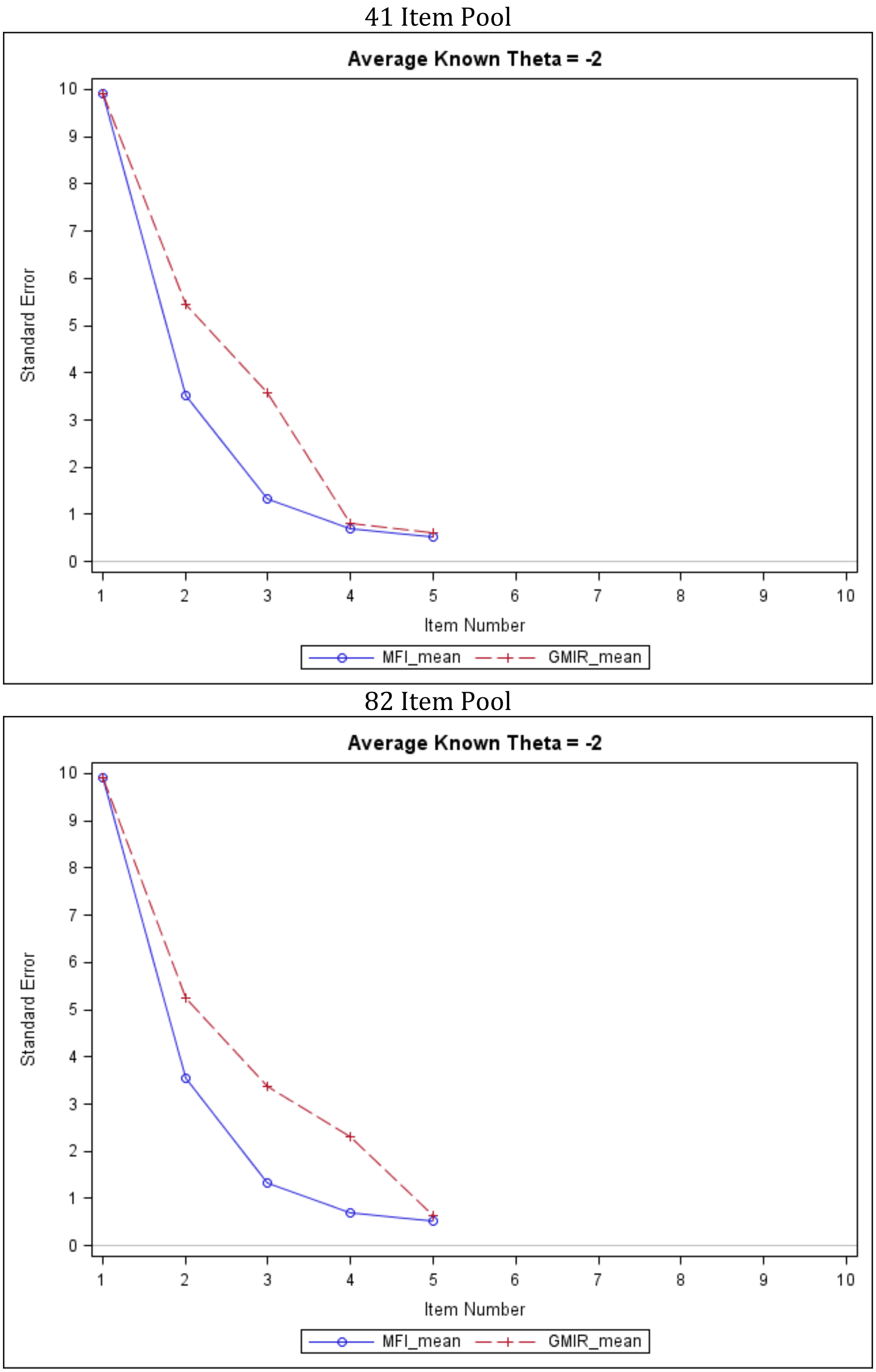


Figure 21C. Plots of Mean RMSE Conditional on Item Number for Known Theta = 2, the 9 Item Stopping Rule, and Negatively Skewed Distributed Populations

In Figures 22-26, conditional plots of SE by item are shown for the following theta groups: -2, -1, 0, 1, and 2. The SE by item plots did not vary between the normally distributed population conditions and the negatively skewed population conditions for any theta value. Only the normal conditions are shown below; the negatively skewed plots and any conditions not shown can be found in the Appendix A. In general, across all theta values, GMIR resulted in larger SE in items 2-4, but MFI and GMIR resulted in similar SE values at the first and last few items. For theta=-2, shown in Figures 22, GMIR conditions resulted in higher SE values at items in the middle of the measure. Using the 41 item pool, values were higher from items 2-3 for the 5 item stopping rule and items 2-4 for the 7 and 9 item stopping rules. Using the 82 item pool, SE values were higher from items 2-5 for the 5 item stopping rule and items 2-6 for the 7 and 9 item stopping rules. SE values were similar for GMIR and MFI at the first and last items for each condition.

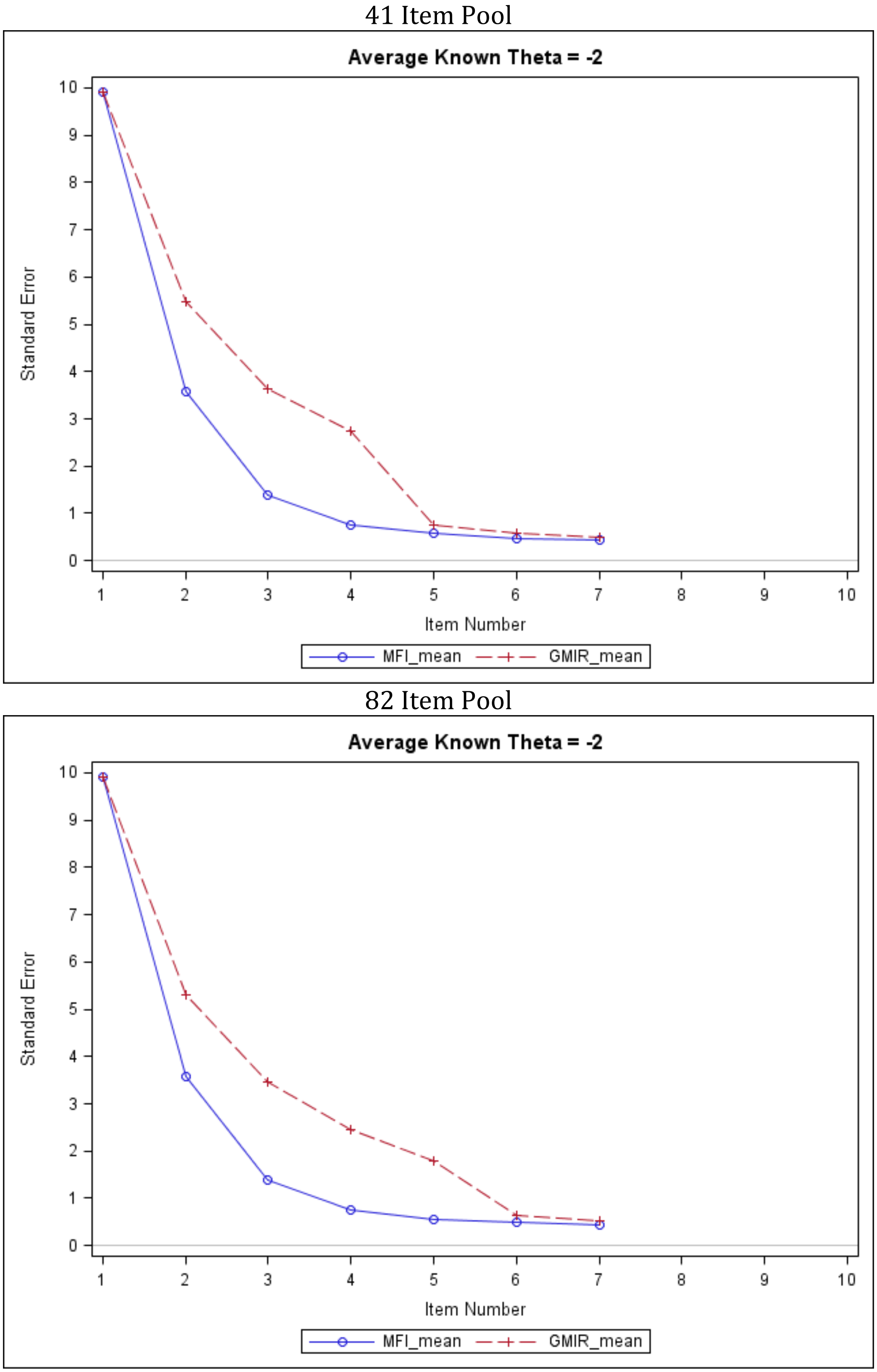
Figure 23 shows the conditional SE values for each test items for theta=-1. Across all conditions, GMIR simulation SE values were higher from items 2 to 4. Since the plots of all 12 conditions are similar, only the 9 item stopping rule/ normally distributed population/ 41 item pool condition is shown in the figure. GMIR and MFI simulations resulted in similar SE values for all other items. Conditional SE values for each item for theta=0 are shown in Figure 24. GMIR resulted in higher SE values at items 2 and 3 across all conditions. Since the plots of all 12 conditions are similar, only the 9 item stopping rule/ normally distributed population/ 41 item pool condition is shown in the figure. For theta=-1, conditional SE values for each item are shown in Figure 25. Slightly higher SE values were found using GMIR than MFI at items 2 and 3. At the first item administered and any items administered after 3, SE values were similar for GMIR and MFI. This pattern was consistent across stopping rule, item pool, and population distribution conditions, so only the 9 item stopping rule/ normally distributed population/ 82 item pool condition is shown in the figure. Conditional SE values for each item are shown in Figure 26 for theta=2. GMIR produced higher SE values at items 2 and 3 for 41 item pool conditions and at items 2-4 for 82 item pool conditions. Since the plots are similar across stopping rules, only the 9 item stopping rule and normally distributed population with the 41 and 82 item pool conditions are shown in the figure.



41 Item Pool

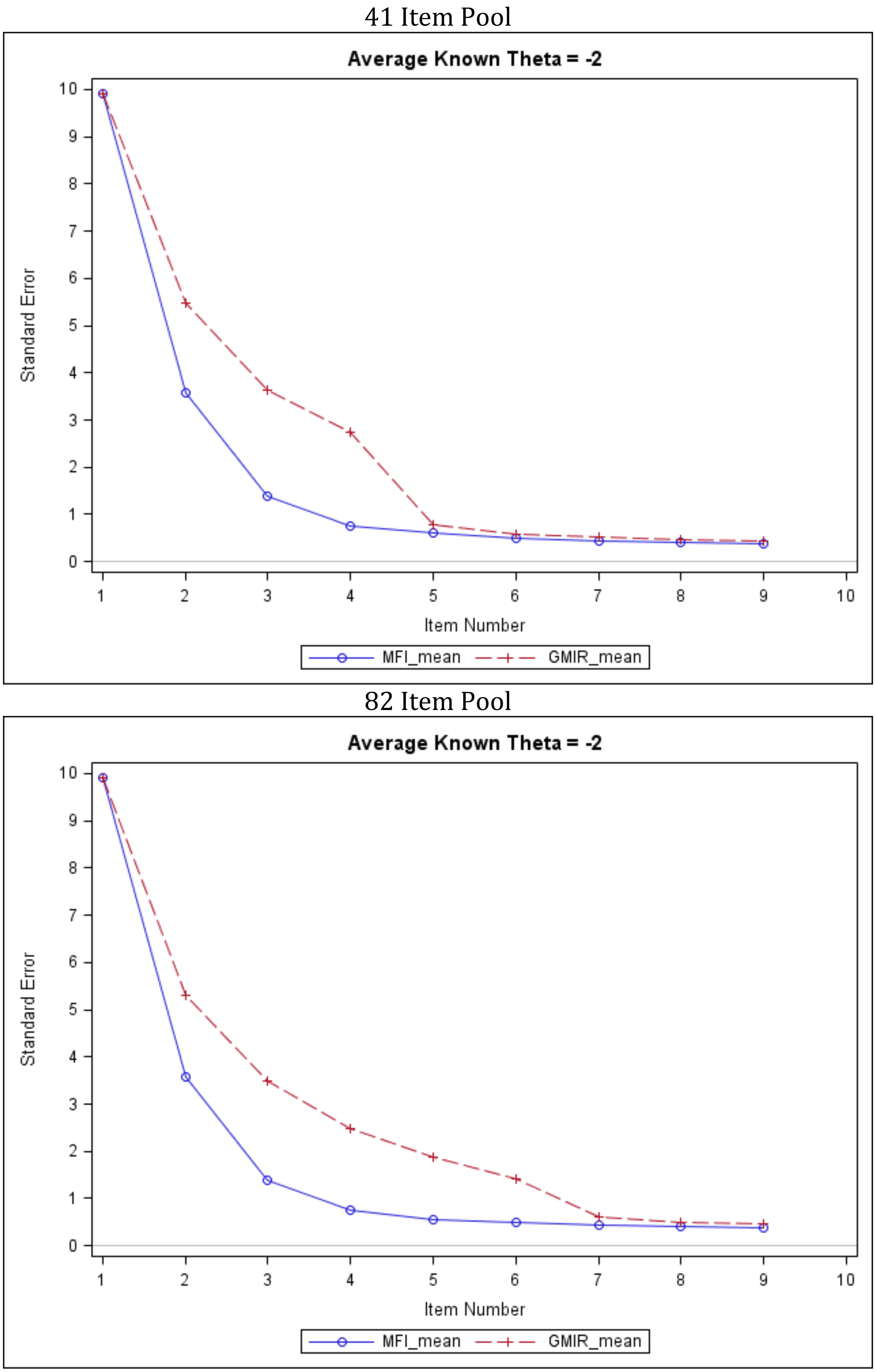
82 Item Pool

Figure 22A. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 5 Item Stopping Rule, and Normally Distributed Populations


41 Item Pool

82 Item Pool

Figure 22B. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 7 Item Stopping Rule, and Normally Distributed Populations



41 Item Pool

82 Item Pool

Figure 22C. Plots of Mean SE Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, and Normally Distributed Populations

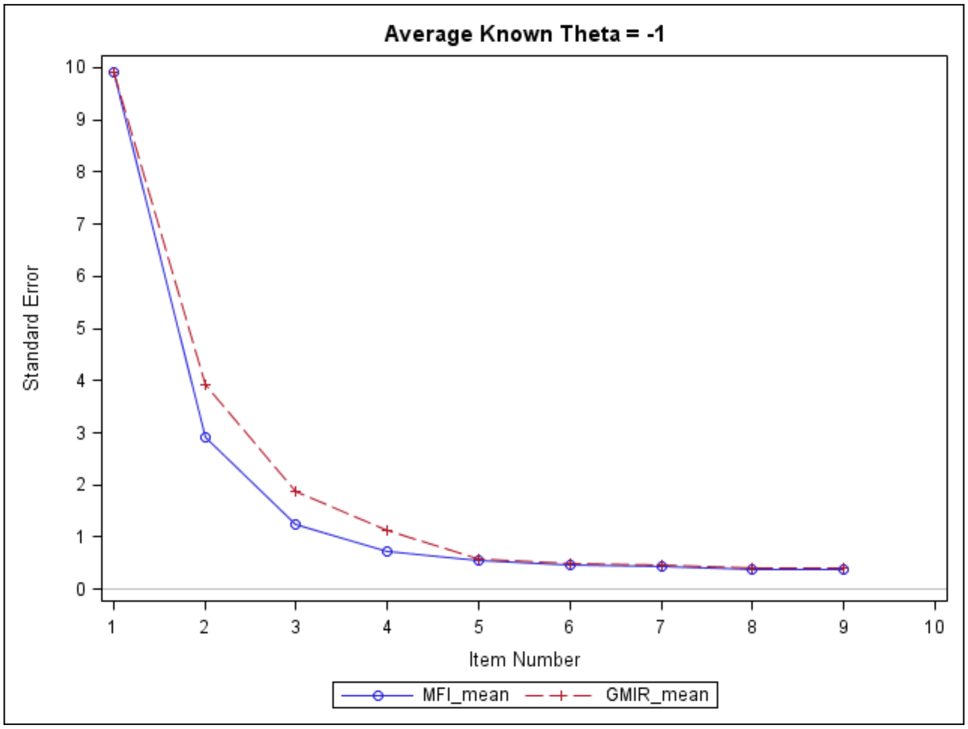


Figure 23. Plot of Mean SE Conditional on Item Number for Known Theta = -1, the 9 Item Stopping Rule, Normally Distributed Population, and 41 Item Pool

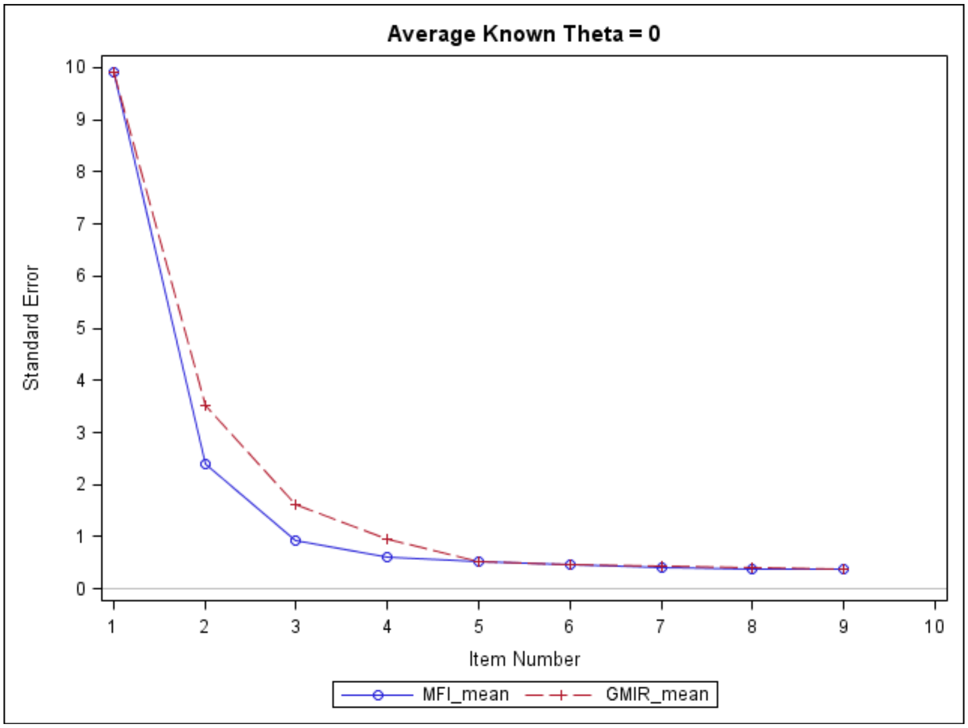


Figure 24. Plot of Mean SE Conditional on Item Number for Known Theta = 0, the 9 Item Stopping Rule, Normally Distributed Population, and 41 Item Pool

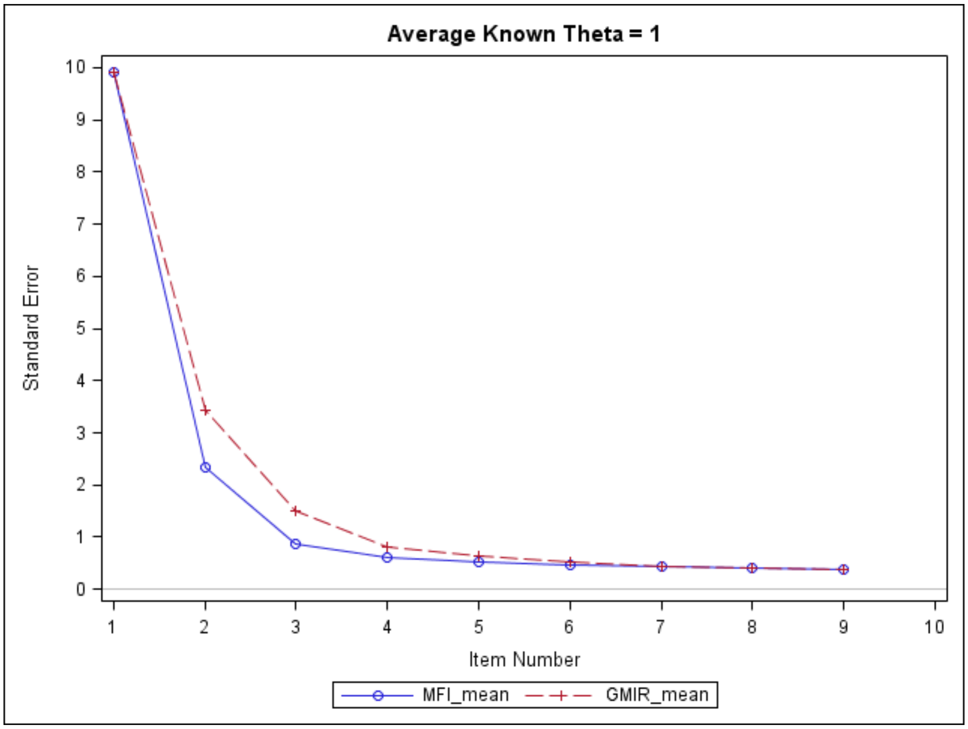


Figure 25. Plot of Mean SE Conditional on Item Number for Known Theta = 1, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool

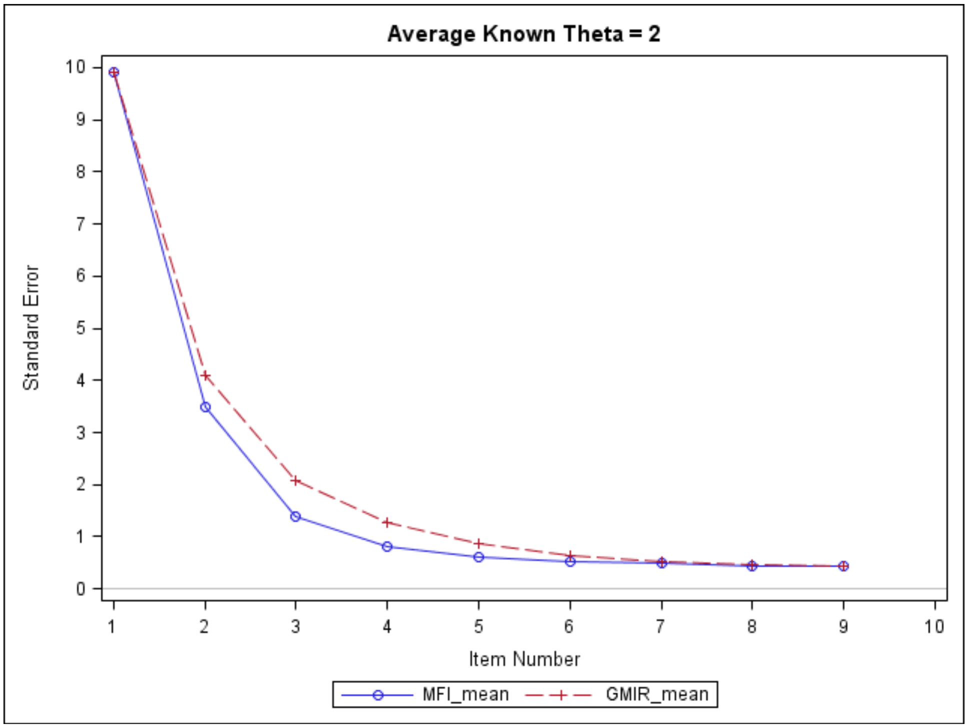


Figure 26. Plot of Mean RMSE Conditional on Item Number for Known Theta = -2, the 9 Item Stopping Rule, Normally Distributed Population, and 82 Item Pool

TEST EFFICIENCY

The mean, minimum, and maximum number of items administered (NIA) for each condition averaged across 100 replications are shown in Table 6. When the stopping rule was 5 items or an SE of .54, GMIR simulations administered 4.858 items on average and MFI simulations administered 4.99 on average. When the stopping rule was 7 items or an SE of .46, GMIR administered 6.488 and MFI administered 6.283. When the stopping rule increased to 9 items or and SE of .40, GMIR conditions administered 8.578 items on average and MFI conditions administered 8.223 items on average. For every condition with the 5 item stopping rule, the minimum and maximum number of items administered were 4 and 5. Likewise, for the 7 item stopping rule conditions, the minimum and maximum number of items administered were 6 and 7. For all the 9 item stopping rule conditions, the minimum and maximum number of items administered were 8 and 9. Using GMIR, the 7 and 9 item stopping conditions, and 41 item pool the number of items administered was slightly greater than using the 82 item pool. The NIA was 8.74 and 8.41 on average for the 9 item stopping rule with the 41 and 82 item pools respectively. The NIA was 6.56 and 6.42 on average for the 7 item stopping rule with the 41 and 82 item pools respectively.

Condition			Number of Items Administered		
			Mean	Min	Max
Normally Distributed Population	41 Item Pool	5 item			
		GMIR	4.87	4	5
		MFI	4.99	4	5
		7 item			
		GMIR	6.57	6	7
		MFI	6.28	6	7
		9 item			
		GMIR	8.74	8	9
		MFI	8.25	8	9
	82 Item Pool	5 item			
		GMIR	4.85	4	5
		MFI	4.99	4	5
		7 item			
		GMIR	6.42	6	7
		MFI	6.28	6	7
		9 item			
		GMIR	8.43	8	9
		MFI	8.19	8	9
Negatively Skewed Population Distribution	41 Item Pool	5 item			
		GMIR	4.86	4	5
		MFI	4.99	4	5
		7 item			
		GMIR	6.55	6	7
		MFI	6.29	6	7
		9 item			
		GMIR	8.75	8	9
		MFI	8.26	8	9
	82 Item Pool	5 item			
		GMIR	4.85	4	5
		MFI	4.99	4	5
		7 item			
		GMIR	6.41	6	7
		MFI	6.28	6	7
		9 item			
		GMIR	8.39	8	9
		MFI	8.19	8	9

Table 6. Number of Items Administered Descriptive Statistics Averaged across 100 Replications

Chapter 5: Discussion

This chapter contains three main sections, which discuss the study results. First, the research questions are addressed based on the findings. Next, the practical applications the conclusions are discussed. Finally, the limitations of the study and future research directions are addressed.

RESEARCH QUESTIONS

How do the MFI and GMIR item selection methods' performances compare for CATs with small numbers of items?

A number of differences were found between the performance of MFI and GMIR item selection methods; some of these differences favored MFI and some favored GMIR. When using MFI, there were three times fewer overall nonconvergent cases as compared to GMIR, with 9.6 and 3.4 total nonconvergent cases on average. This finding is inconsistent with the findings of Chang and Dodd (2013) who found more conconvergent case with MFI than with GMIR. While the number of nonconvergent cases using GMIR was consistent with Chang and Dodd's results, they found over ten times as many nonconvergent cases with MFI than found in the present study. This difference is probably due to fewer items being administered in the current study; 20 items were administered in the Chang and Dodd(2012) study. While the cases not reaching MLE were more similar, there were 5 times as many out of range cases with GMIR than with MFI. Both MFI and GMIR simulations resulted in mean θ estimates that were close to 0. GMIR was less consistent, resulting in a wider range of mean θ estimates, but had a mean final θ estimate closer to 0 in the 5 item condition. The standard deviations of the θ estimates were both close to 1 and the SE values were comparable using MFI and GMIR. Mean θ estimates close to 0 and SD close to 1 for both GMIR and MFI is consistent with previous research (Chang and Dodd, 2013; Han, 2009; Han, 2010).

Measurement precision was better using GMIR when the 5 item stopping rule was used and fewer items were administered. Mean correlation coefficients were 0.875 and 0.859, mean bias values were 0.029 and 0.077, and mean RMSE values were 0.554 and 0.595 for GMIR and MFI respectively. At the 7 and 9 item stopping rules, the measurement precision values were comparable between the two item selection methods. These results are consistent with and expand on the findings of previous studies (Chang and Dodd, 2013; Han, 2009). Chang and Dodd found GMIR had slightly smaller mean SE, mean bias, and mean RMSE at the early stages of CAT and when the CATs were shorter, across an NIA range of 12-20 items. Interestingly, this study did not find GMIR outperforming MFI in the 7 and 9 item conditions, which are objectively short tests even though they were the medium and long test conditions in this study. Han (2009) found GMIR resulted in a slightly smaller SE compared to MFI.

To further investigate the measurement precision of the two item selection methods, MFI and GMIR were compared across known θ values. Item selection method performance varied across the θ scale and by outcome measure. At the extreme negative end of the θ scale, at θ=-3.5, GMIR resulted in a larger bias and SE, but MFI resulted in a larger RMSE under certain conditions (5 or 9 item/normal/41 pool). When -3.5< θ<-1, GMIR resulted in larger RMSE and SE values. Around θ=0, in the middle of the scale, MFI resulted in a larger RMSE and bias, especially in the 5 item conditions. On the positive end of the θ scale, GMIR resulted in larger RMSE and SE values in some of the conditions, for other conditions the item selection methods performed comparably. These results are inconsistent with previous studies. This study found GMIR bias was larger at the extreme negative θ, while Chang and Dodd (2013) found GMIR was smaller at the extreme θ values and comparable across the rest of the θ scale. Han (2009) found GMIR and MFI measurement precision was comparable across the θ scale. This inconsistency could be due to the number of items administered. Chang and Dodd used a stopping rule that resulted 18-19 items on average and Han used a 40 item stopping rule. In the current study, most differences across the θ scale were larger at the 5 item condition, decreasing as more items were administered.

Comparing MFI and GMIR across item number, generally MFI results in better measurement precision at the beginning of the test and GMIR results in better measurement precision towards the end of the test. Across θ groups, using MFI, bias was smaller in the first few items, but slightly larger in the last item, especially in the 7th and 9th item. Across all five θ groups, MFI SE was smaller from around items 2 to 4 and then SE values were comparable for the remainder of the test items. RMSE values varied across items more across θ groups; no clear pattern of performance emerged that favored GMIR or MFI. This finding is inconsistent with the previous findings of Chang and Dodd (2013) who found GMIR simulations to have smaller SE, bias and RMSE at the early stages of CAT. Chang and Dodd found these differences in measurement precision were largest at θ=2, where in this study, differences were smallest at θ=2. This could be due to the shape of the item pool information. When compared to the item pool used in Chang and Dodd (Davis, 2004), the item pool used for this study had similar information at θ=2, but provides less information along other parts of the θ scale.

How do the item pool size, mismatch of item pool and population latent trait distributions, and test length affect the measurement precision?

The item pool size was found to have minimal affect on the outcome variables reported. This result is consistent with the findings that item pool size can be as small as 30 items (Dodd, 1990; Dodd & de Ayala, 1994). No difference was found in the number of nonconvergent cases, descriptive statistics of the final θ estimates, standard errors, correlations coefficients, or mean RMSE values. Conditions using GMIR and MFI resulted in slightly higher mean bias values using the 82 item pool than with the 41 item pool, especially in the 5 item condition, but these differences were too small to have practical importance. When using the GMIR item selection procedure and the 7 and 9 item stopping rules, the NIA was slightly larger, by .33 and .14 items, using the smaller item pool than with the larger item pool. These differences are too small to have practical importance.

The mismatch of item pool and population latent trait distribution was also found to have minimal affect on the outcome variables. No differences were found in descriptive statistics of the mean θ estimates, mean standard errors, mean bias, mean RMSE, mean correlations, or mean NIA. The mismatch of item pool and populations distributions was found to slightly affect the number of nonconvergent cases. The negatively skewed population conditions resulted in one more out of range case and one more case that did not reach MLE than the normally distributed population when using GMIR. These finding are consistent with Lee and Dodd (2012) who found that the measurement precision of a polytomous CAT using the PCM was relatively robust to the mismatch between item pool and trait distribution.

The test length was found to affect number of nonconvergent cases, the mean final θ estimates and mean standard deviations, and the measurement precision. The number of nonconvergent cases not reaching MLE was four times greater with the 5 item condition that with the 7 or 9 item conditions. Similarly, using MFI, the mean final θ estimates were closer to zero with the 7 and 9 item conditions than with the 5 item conditions, -0.07 as compared to -0.02. The mean standard deviations were slightly closer to 1 as the number of item increased. The correlation coefficients, mean bias, and mean RMSE values all showed better measurement precision when more items were administered. When MFI was used, the mean bias was higher using the 5 item stopping rule than with the 7 and 9 item stopping rule, 0.077 as compared to 0.023. While the 5 item stopping rule correlation of 0.867 is a high positive correlation, when the 7 or 9 item stopping rules were used, correlation coefficients were greater than or equal to .9, which constitutes a very high positive correlations (Hinkle, Wiersma, & Jurs, 2003). RMSE decreased from the 5 item to 7 item to the 9 item stopping rule, with values of 0.58, 0.48 and 0.42. Looking at the plots across θ values, MFI resulted in a larger bias and RMSE values around θ=0, especially with the 5 item stopping rule. The difference in RMSE was 0.1, 0.5, and 0.35 in the 5 item, 7 item, and 9 item stopping rule conditions and the difference in bias was 0.1 in the 5 item stopping rule condition. Item selection method performance in terms of test efficiency varied according to test length. GMIR administered slightly fewer items on average when the 5 item stopping rule was used, 4.86 as compared to 4.99 with MFI. However, using the 7 and 9 item stopping rules, MFI outperformed GMIR in terms of test efficiency. MFI administered 6.28 and 8.22 items on average as compared to 6.49 and 8.58 items on average.

How do interactions among item pool size, mismatch of item pool and population latent trait distributions, and test length affect the MFI and GMIR item selection methods' performance?

A few interactions among the item pool size, mismatch of item pool and population distribution, and test length were found to affect the item selection method's performance in terms of nonconvergence, NIA, and plots of measurement precision across θ values. The number of nonconvergent cases varied by item pool size, test length, and item selection method. Using GMIR item selection method, there were more nonconvergent cases when the 5 item test was combined with the larger item pool. There were 4 more cases not reaching MLE in the 5 item stopping rule and 82 item pool condition than the 5 item stopping rule and 41 item pool condition. However, using MFI, there was one more case of nonconvergence when the 5 item stopping rule was combined with the 41 item pool condition versus the 82 item pool condition. Using the GMIR item selection method, the number of items administered was slightly higher by .1 and .3 items in the 7 and 9 item conditions when the 41item pool was used versus the 82 item pool. Using MFI or when using GMIR in the 5 item stopping rule condition this was not the case, NIA was similar across item pool sizes.

The plots of RMSE and SE across known θ values showed interaction effects at specific θ values. At θ=-3.5, GMIR resulted in smaller RMSE values by 0.1 than MFI when the population was a match to the item pool and the 5 or 9 item stopping rule and the 41 item pool was used. This difference was due to MFI performing worse using the smaller item pool, 5 and 9 item stopping rules, and GMIR performing better when the population matched the item pool distribution using the 41 item pool. MFI resulted in smaller SE values by .03 to .15 when θ was less than -1 than GMIR. The difference in performance between the item selection procedures was because GMIR resulted in higher SE values when the 82 item pool and 5 or 9 item stopping rules were used.

While interactions were found for two individual outcome variables, no clear pattern of interaction emerged across variables and item selection methods. This study did not find the interaction found in Keng (2008). Keng found increase bias and RMSE when the item pool was smaller, a mismatch occurred between item pool and population, and the test was shorter. Here the maximum bias and RMSE values were found in the shortest test length and mismatch between population and item pool distributions, but these values were maximized when the larger item pool was used. Also, the values are comparable to the condition when the population and item pool distributions are matching.

PRACTICAL APPLICATIONS

Recently, in the health care and medical fields, computer-based PRO measures have become increasingly more common. With the NIH's investment in and development of PROMIS, computer adaptive testing is at the forefront of modern PRO measurement in clinical research and medical practices. A number of the advantages of CAT make it especially beneficial for PRO measures and understanding the implications of various CAT components within the orientation and constraints of PRO measures is important to the health care field.

This study found that the smaller item pool with 41 items performed as well as the 82 item pool. When item pool development is difficult and requires many resources, the ability to use a smaller pool of items is beneficial (Reeve, 2006). Since the median item pool for PRO measure is 50 items, this result that 41 items is sufficient indicates many current PRO item pools will perform as well as a larger pool (Walker et al., 2010).

Consistent with previous research, the results of this study indicate that a mismatch between the item pool and population latent trait distribution does not negatively affect the measurement precision. In health care, patients may return for numerous treatments and repeated measures of functioning as the level of illness or functioning progresses. As this population distribution shifts, it is important that the PRO measure continues to adequately assess the patients. This finding is encouraging to health care providers looking to continue to precisely measure their patients across numerous treatments.

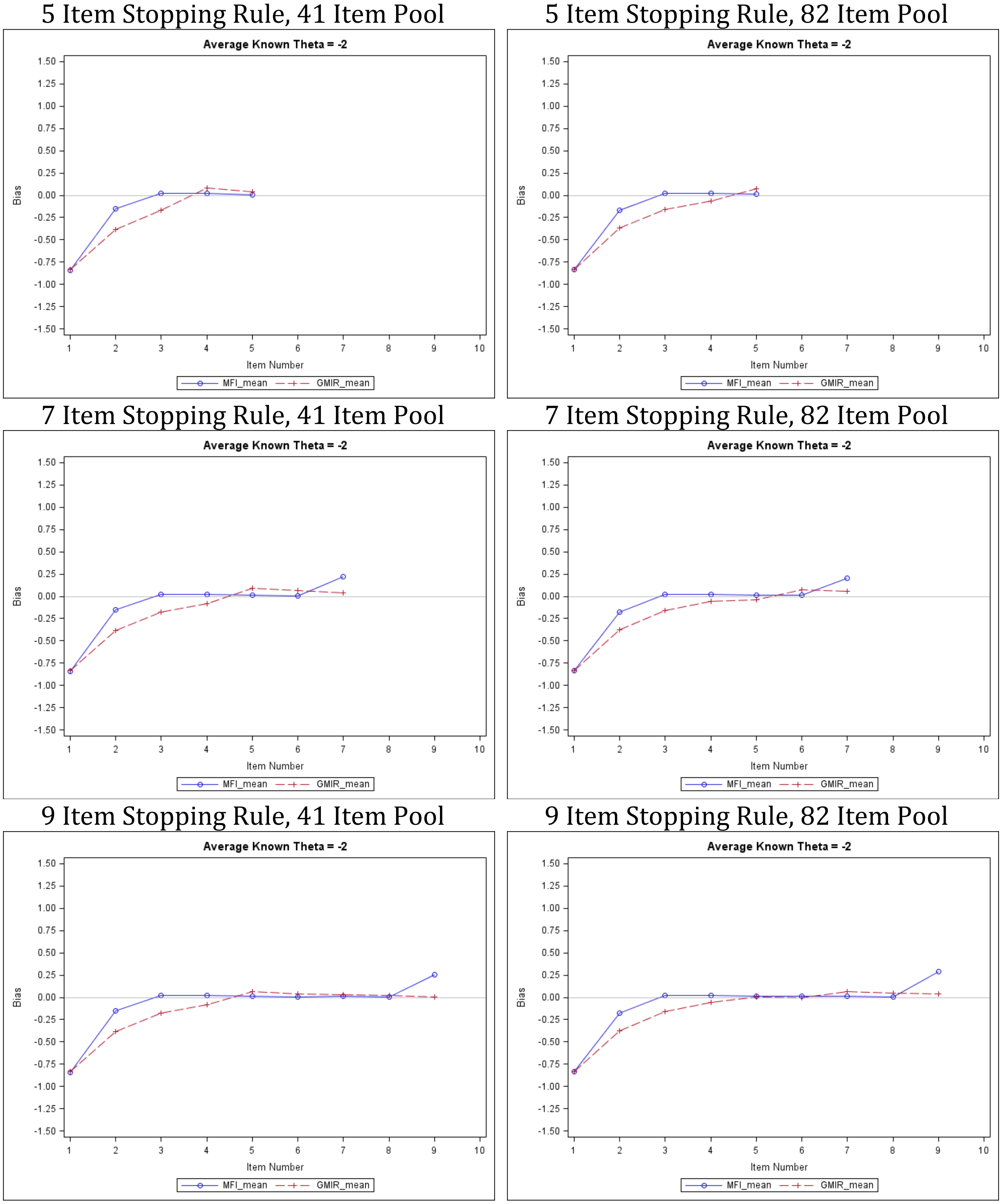
This study demonstrates the varying levels of measurement precision and item selection method performance across test length. Ware and colleagues (2005) suggested that test length be investigated beyond the 5 and 10 item. This study supported the concept that a greater number of item administered results in more precise measurement. However, all three stopping rules resulted in acceptable measurement precision. The mean SE for the 5 item stopping rule when using GMIR, 0.529, is close to the SE used as the stopping rule in some previous PRO studies (Cook et al., 2007; Cook et al., 2008). The results of this study suggest that the 5 item stopping rule is acceptable, but increasing the item number to 7 resulted in greater measurement precision. However, the benefits of administering more items may decrease after 7 items. While the bias of the 7 item stopping rule was an improvement over the 5 item rule, it was comparable to the 9 item stopping rule. There was less decrease in RMSE from 7 to 9 than from 5 to 7 and both of the correlation values for the 7 item and 9 item stopping rules were at or above 0 .90. In terms of item selection method, GMIR displayed better measurement precision than MFI when only 5 items were administered. In future studies, research, or provider practice, using GMIR as opposed to MFI when 5 items are administered could improve the precision of patient assessment. This study did not show an advantage of GMIR over MFI when 7 or more items were administered, so MFI might be preferable for ease of use.

LIMITATIONS AND FUTURE RESEARCH

The findings of this study indicate that the smaller item pool of 41 items was sufficiently large. Conclusions are limited to the conditions of the current study. Investigation into different PRO item pools and different PRO item pool sizes would be beneficial to the health care field. While Walker and colleagues (2010) found a median item pool of 50, they also found a range of item pools from 12 to 282. With the current PROMIS initiative, item pools are being created and adapted from static measures. More information about the optimal item pool size could allow for fewer resources to be directed to this task.

Combination stopping rules of 5 items or .54 SE, 7 items or .46 SE, and 9 items or .40 SE were used for this study. While these stopping rules illustrate the performance of the CAT under these conditions, study of additional test lengths could inform the PRO development. When so few items are administered, one more item can meaningfully decrease the measurement error. Future research should investigate the performance of a CAT using 6 and 8 item stopping rules. The current study used three different SE values in combination with item length for stopping rules, but these SE were used uniformly across examinees of different θ levels. In medical outcome research, a SE value that is conditional on the θ level has been used to allow different precision across the range of θ estimates (Ware et al., 2000; Ware et al., 2005; Ware et al., 2003). Further investigation into the performance of the CAT using variable SE stopping rules would benefit the PRO research field. Additionally, this study only used MLE to estimate the examinee trait levels since a combination stopping rule was used. If PRO studies employ a fixed-item stopping rule for a specific measure or to test different item lengths, the use of WLE in combination with the fixed-length could be beneficial. Previous studies have shown reduced bias and SE in fixed-length CATs (Boyd et al., 2010, Wang et al., 1999; Warm, 1989).

APPENDIX A: PLOTS OF BIAS, RMSE, AND SE CONDITIONAL ON ITEM NUMBER



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

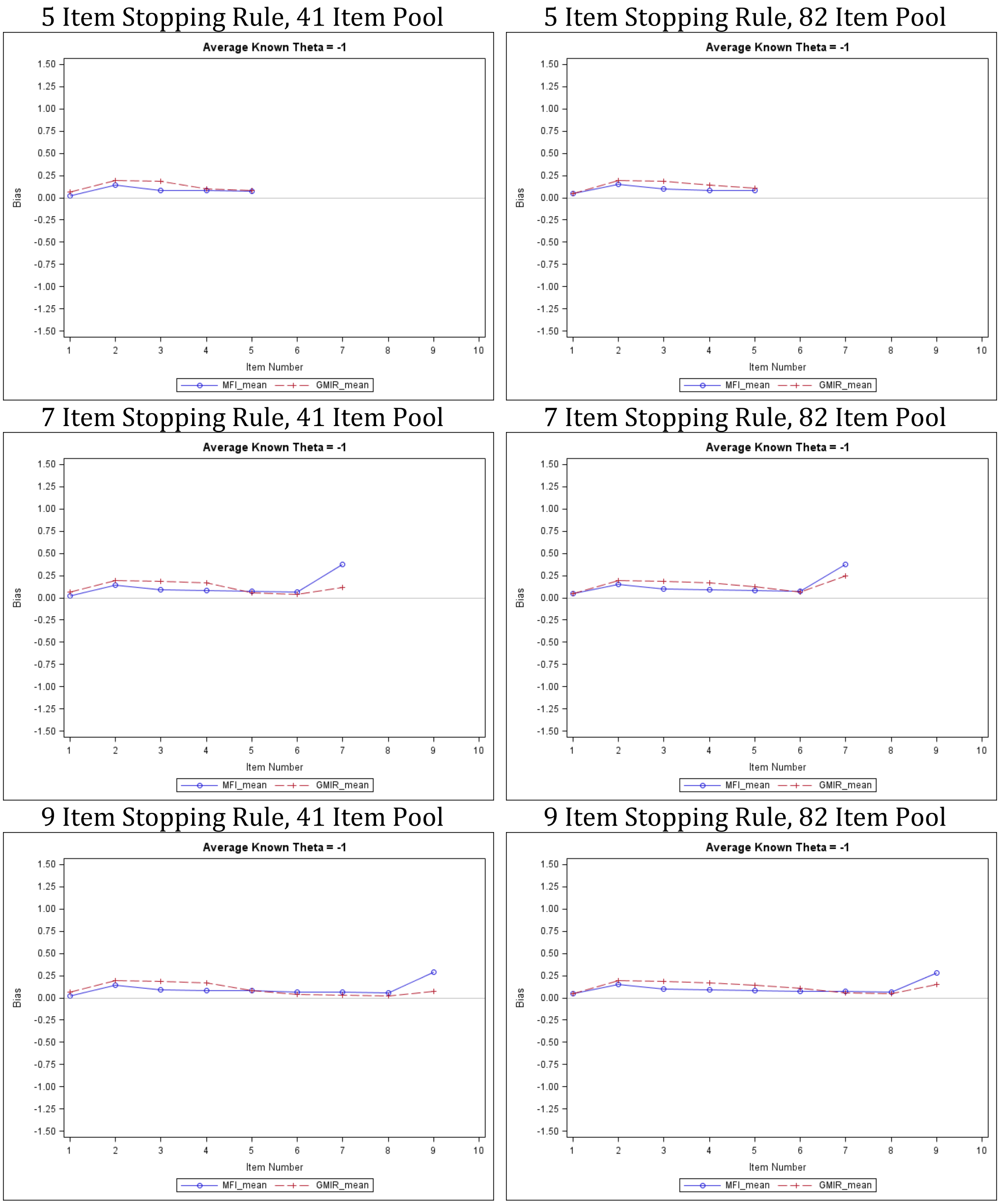
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.1. Plots of Mean Bias Conditional on Item Number for Known Theta = -2 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

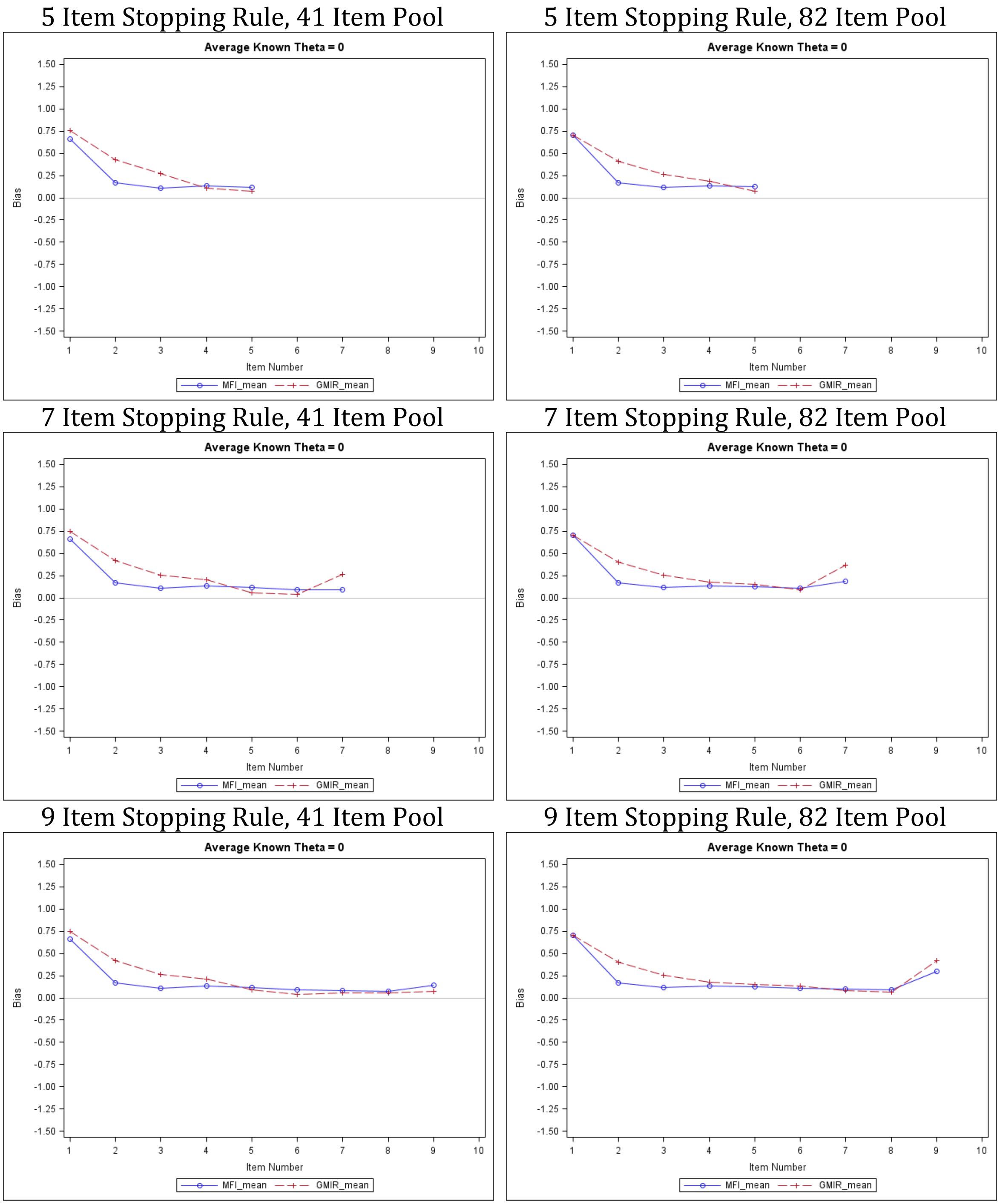
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.2. Plots of Mean Bias Conditional on Item Number for Known Theta = -1 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

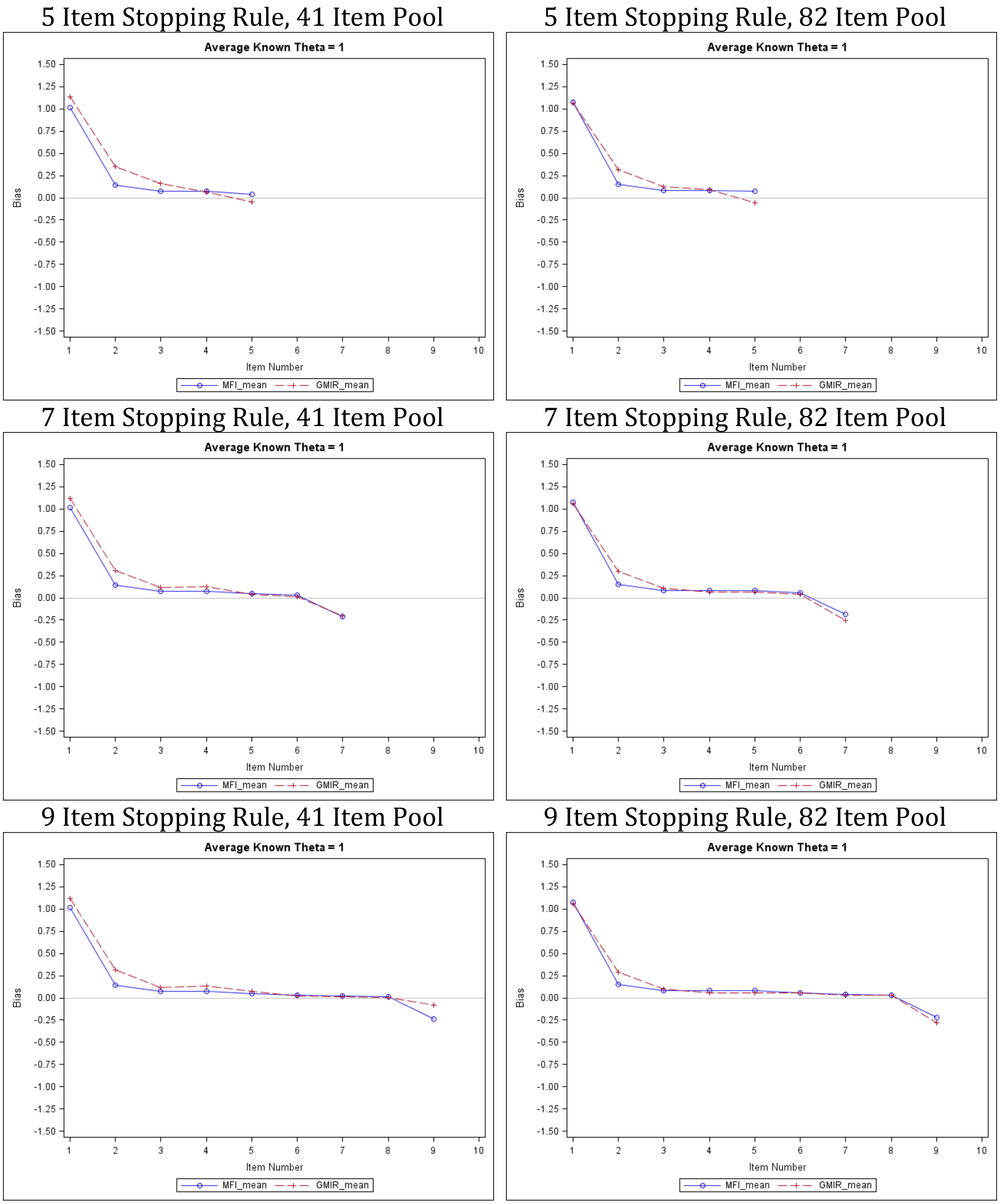
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.3. Plots of Mean Bias Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

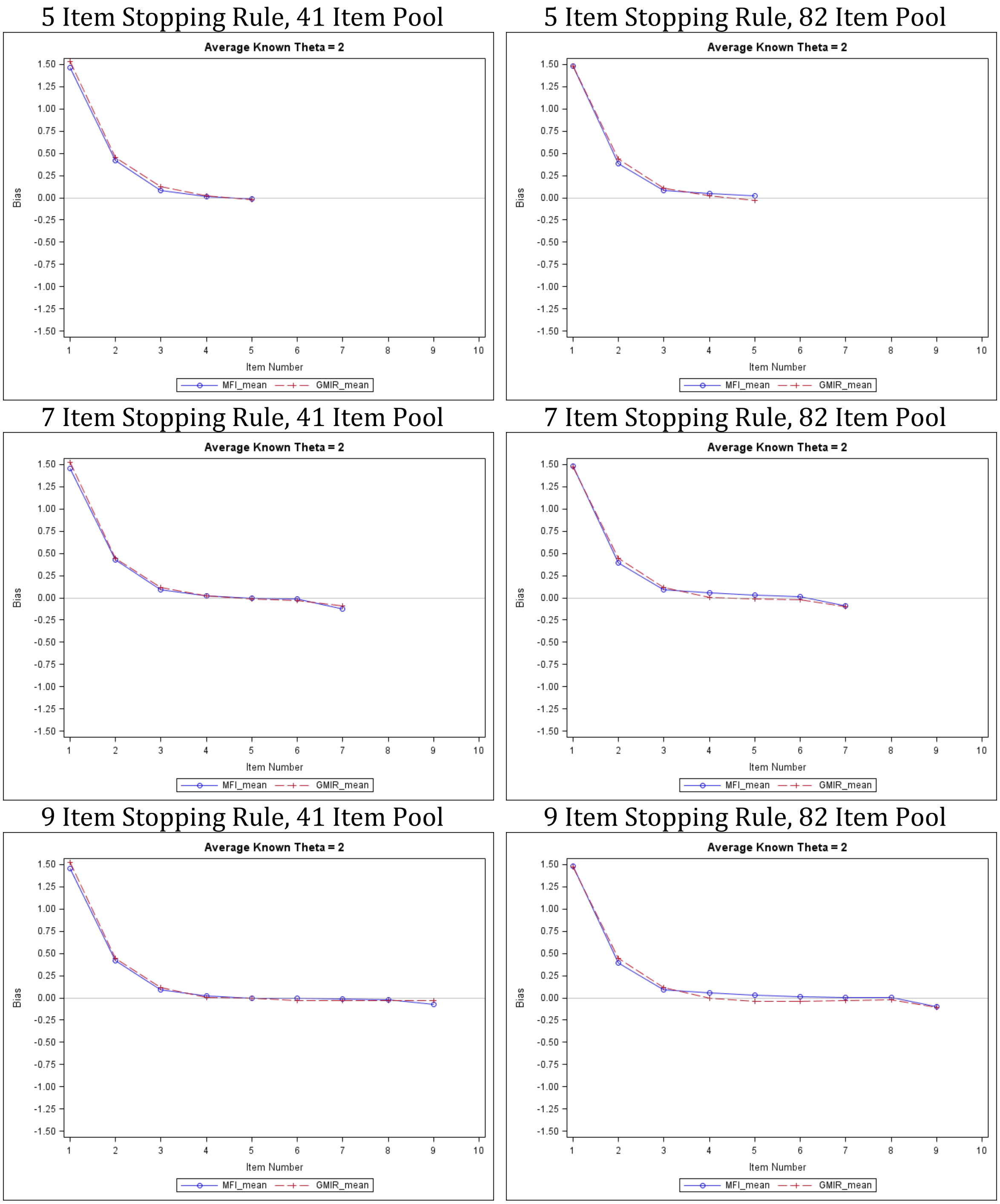
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.4. Plots of Mean Bias Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

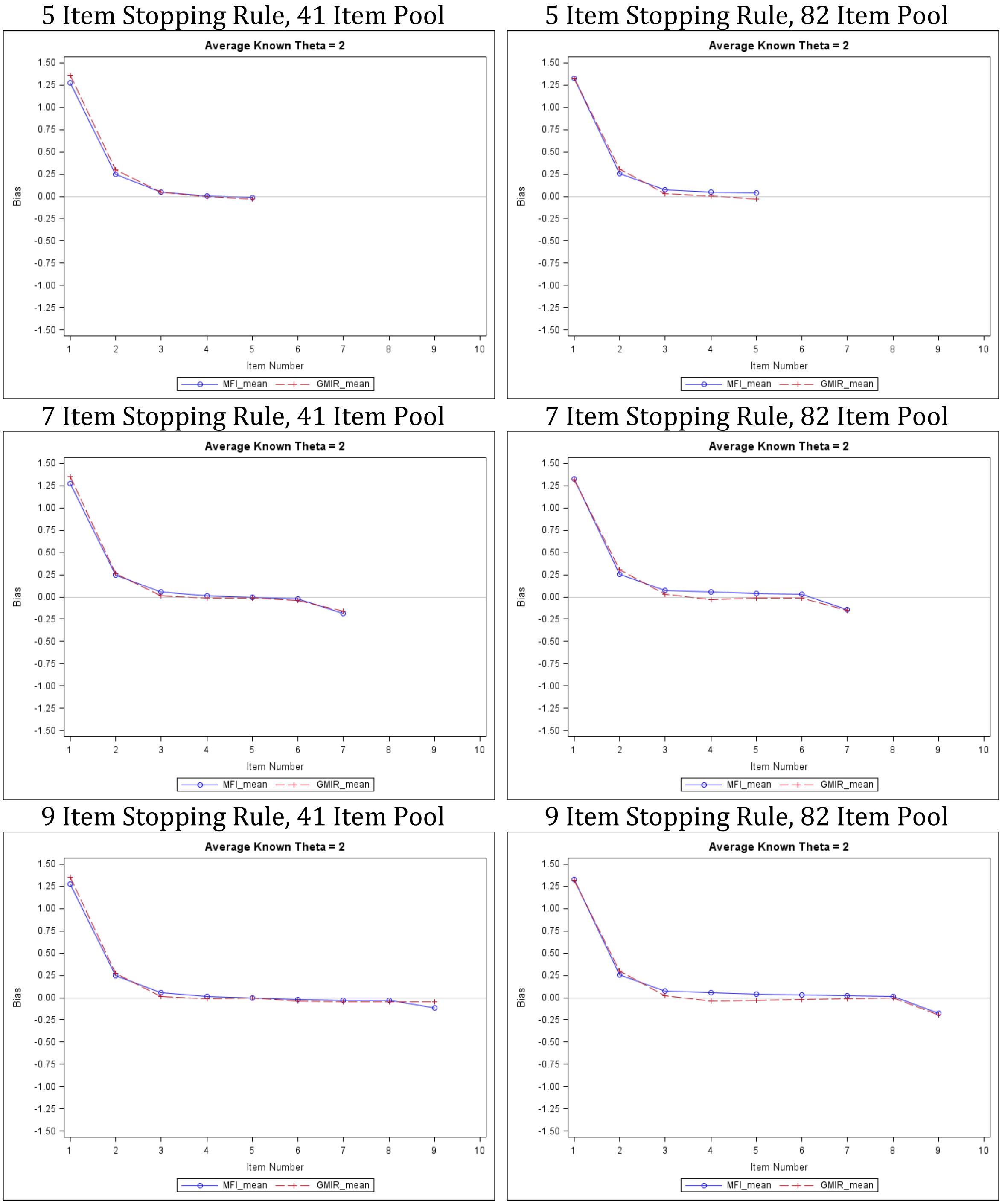
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.5. Plots of Mean Bias Conditional on Item Number for Known Theta = 2 for Normally Distributed Population



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

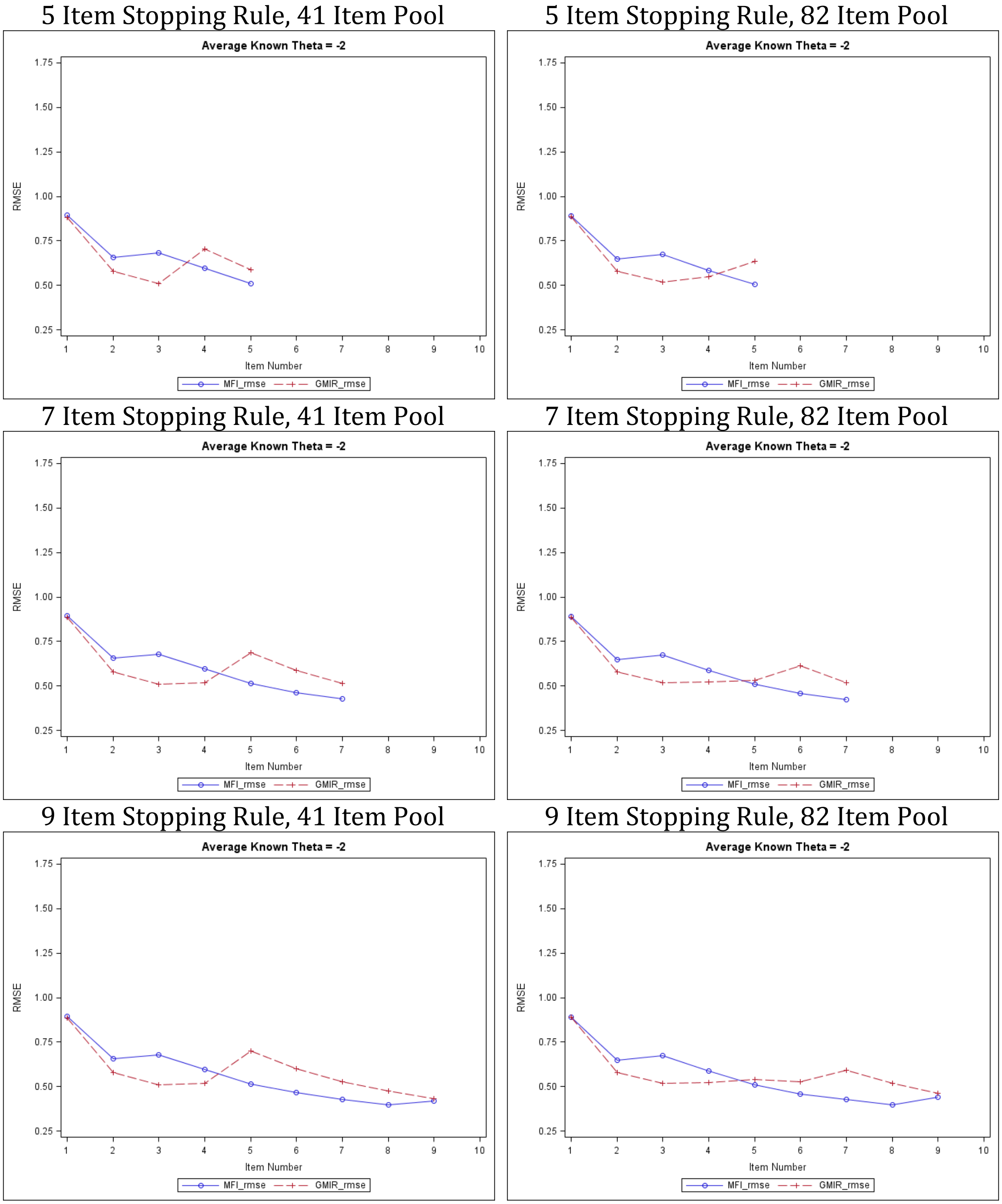
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.6. Plots of Mean Bias Conditional on Item Number for Known Theta = 2 for Negatively Skewed Population Distributions


5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

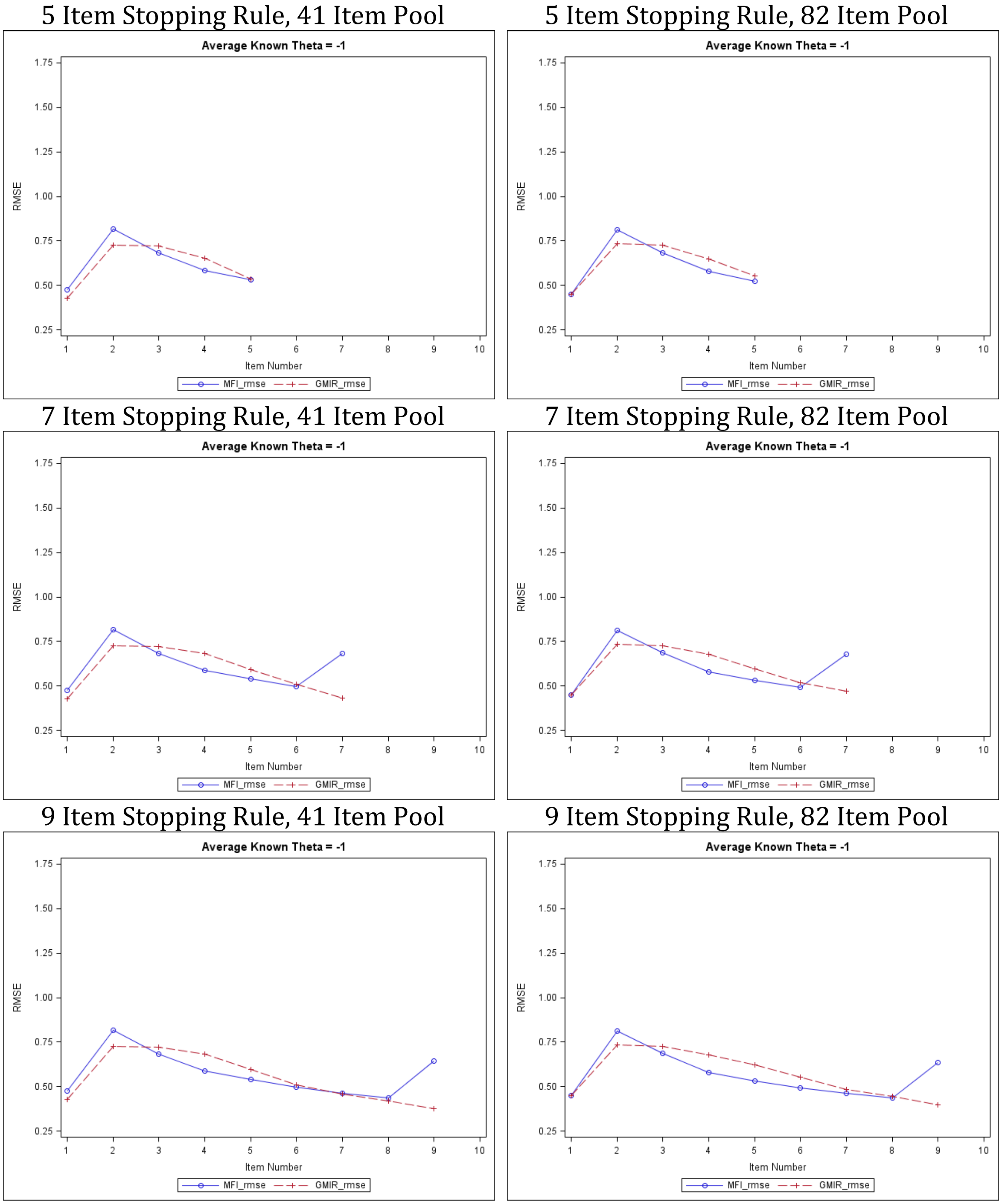
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.7. Plots of Mean RMSE Conditional on Item Number for Known Theta = -2 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

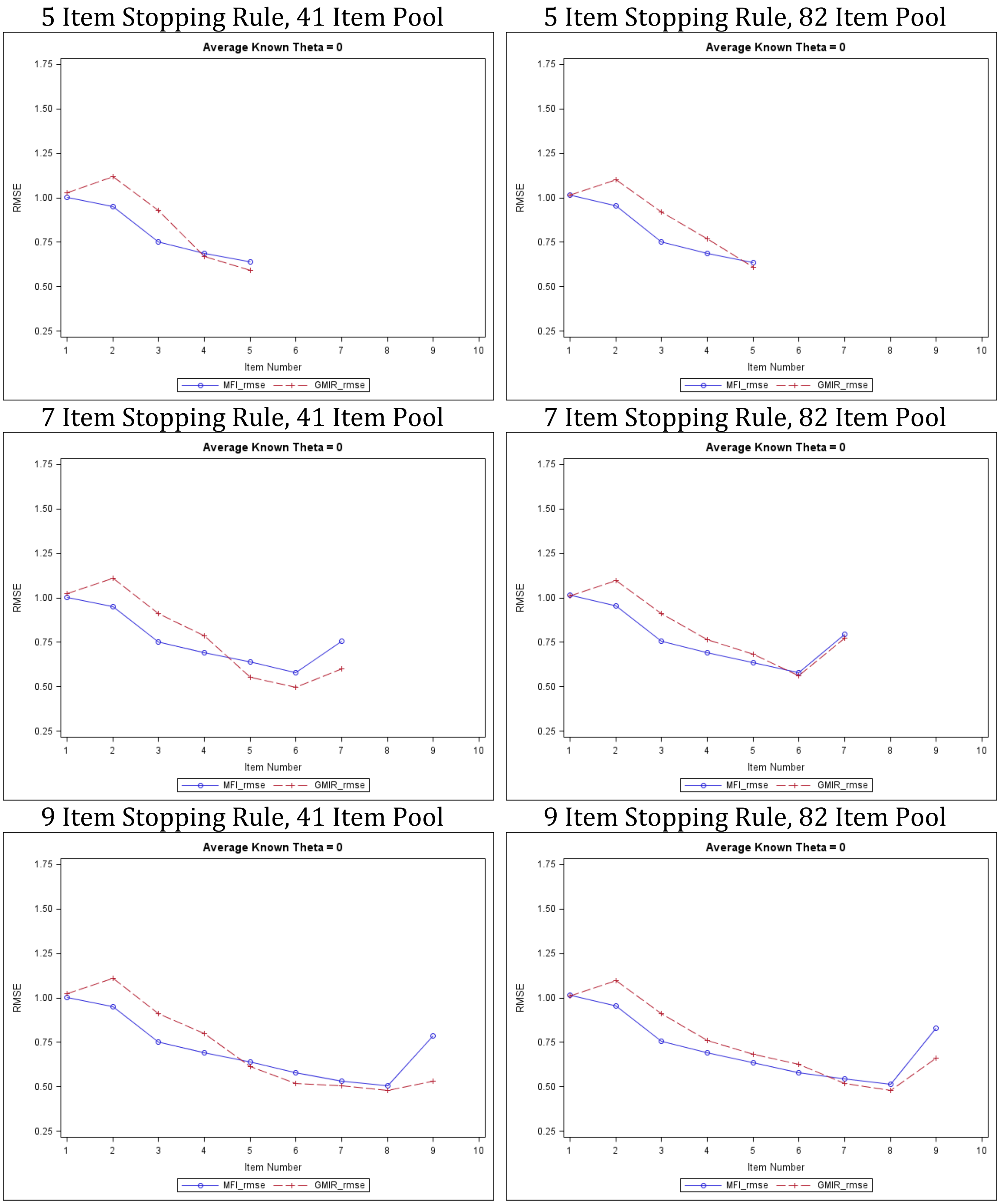
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.8. Plots of Mean RMSE Conditional on Item Number for Known Theta = -1 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

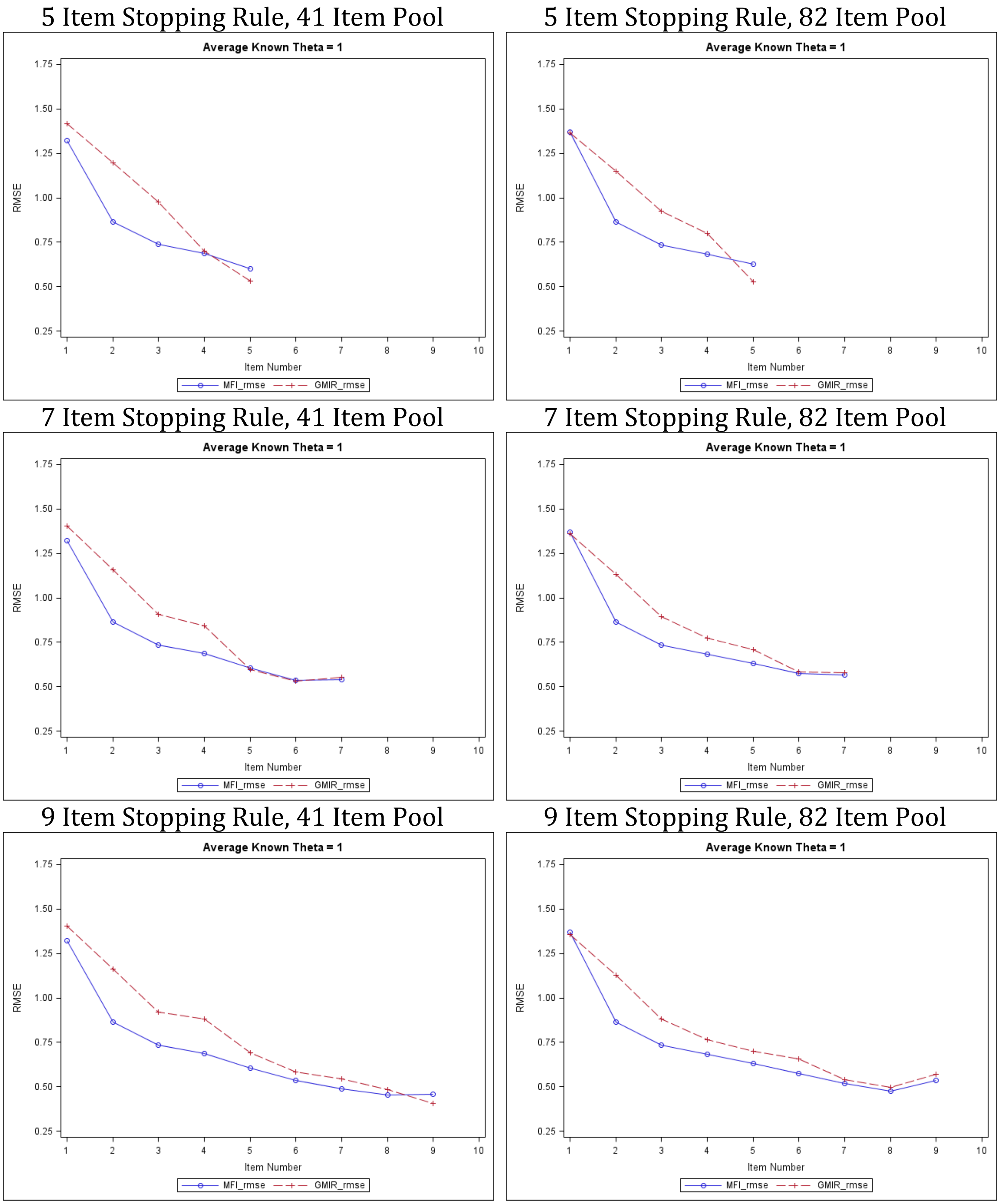
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.9. Plots of Mean RMSE Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distributions


5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

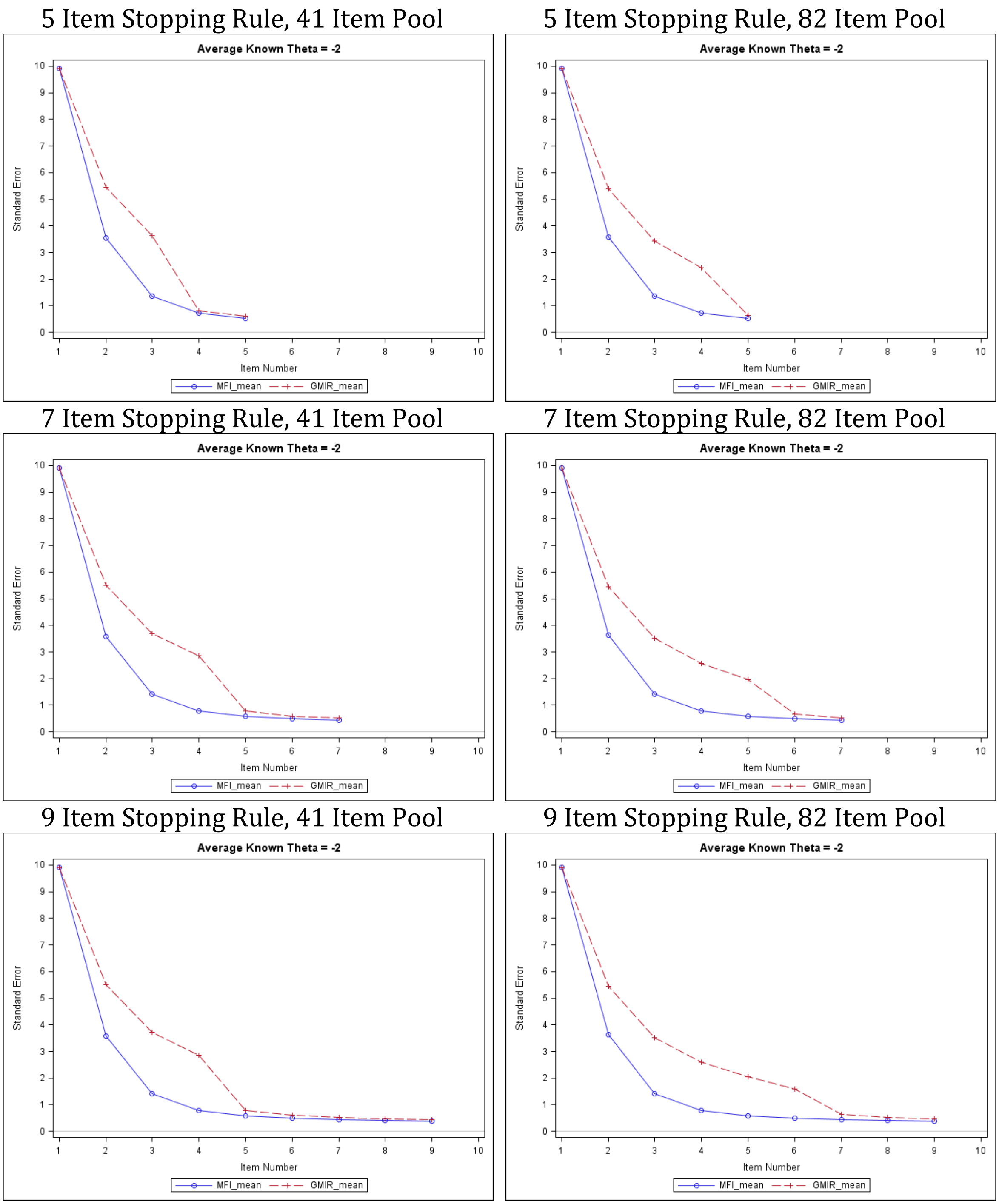
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.10. Plots of Mean RMSE Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distributions


5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

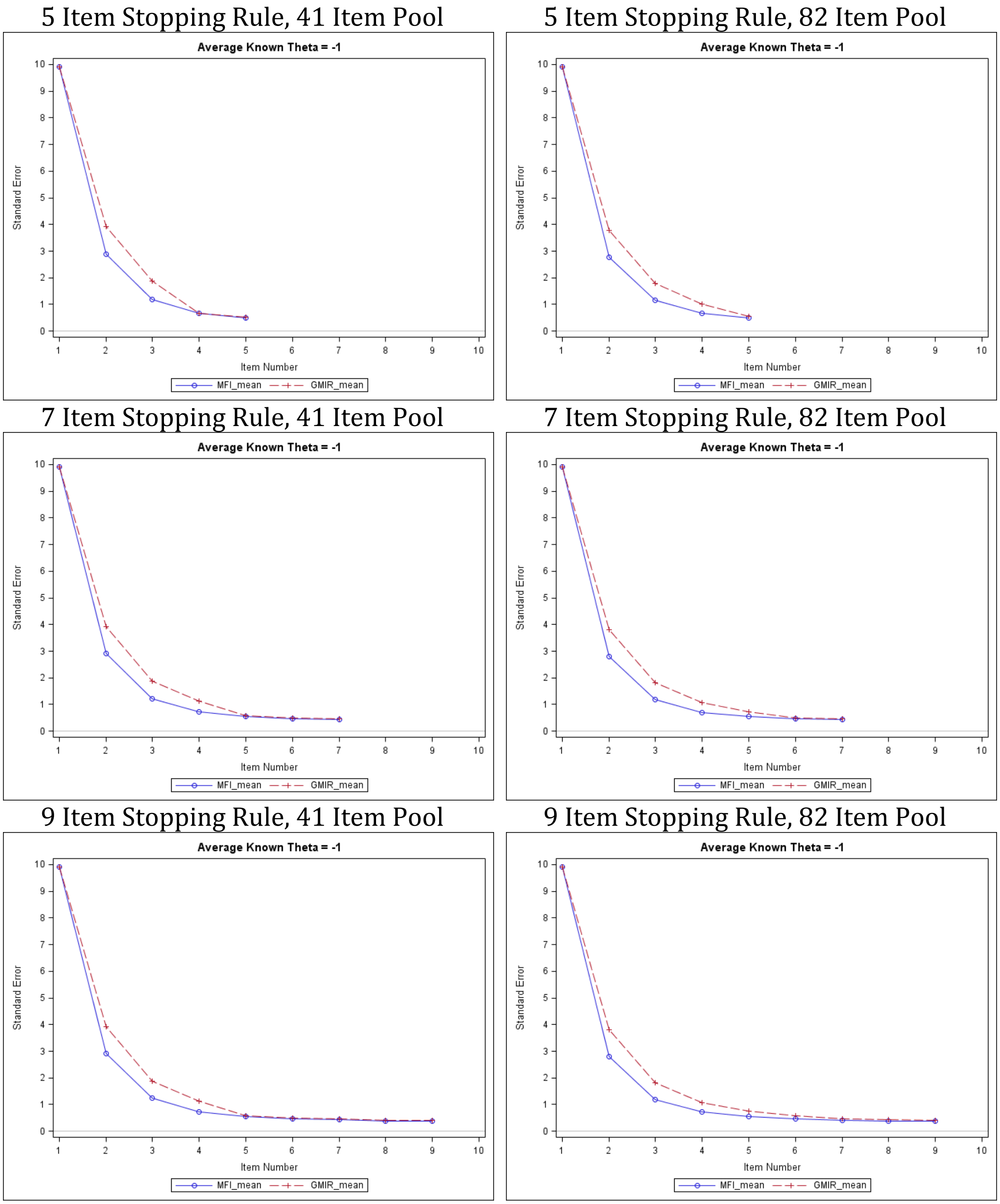
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.11. Plots of Mean SE Conditional on Item Number for Known Theta = -2 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

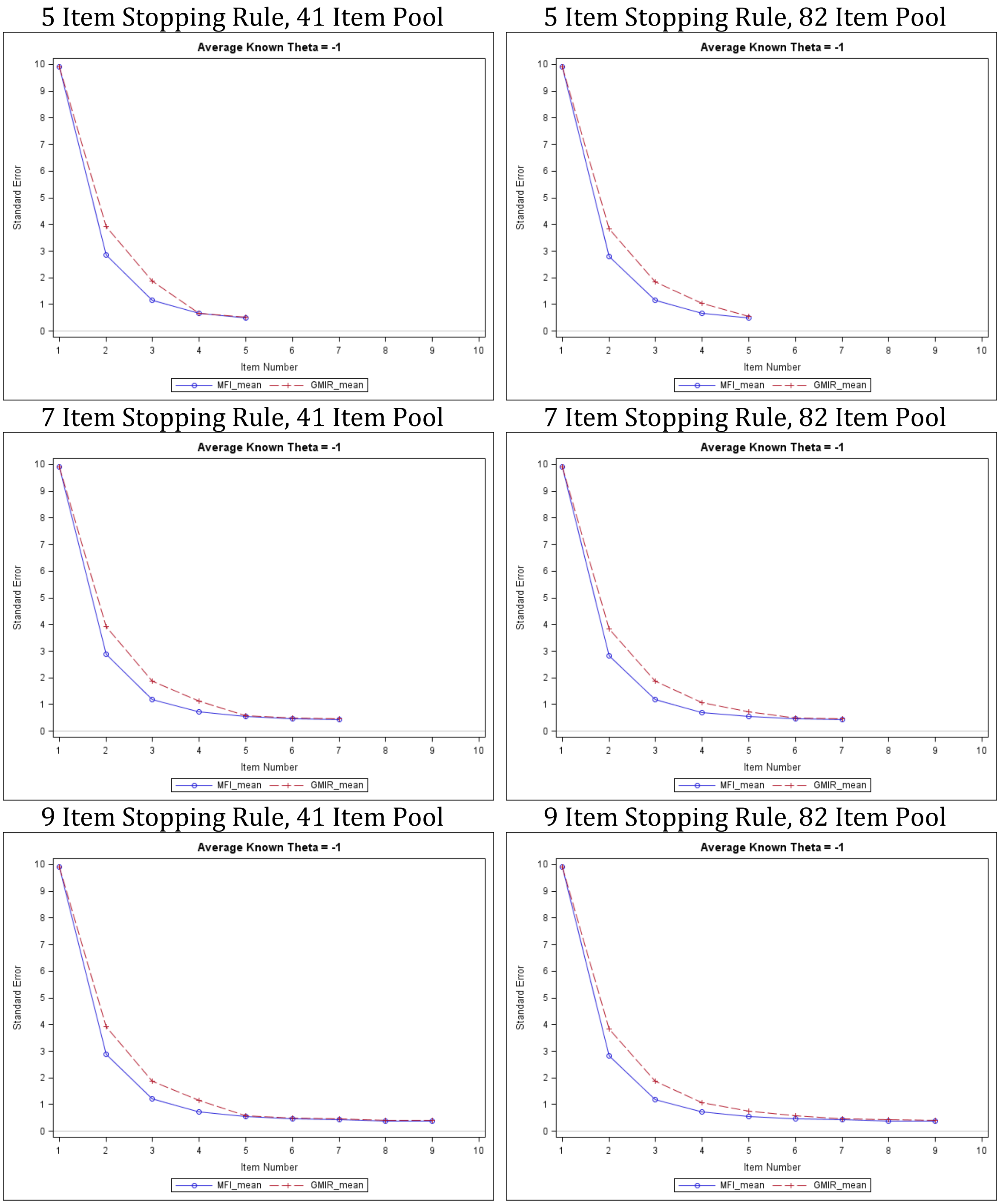
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.12. Plots of Mean SE Conditional on Item Number for Known Theta = -1 for Normally Distributed Populations



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

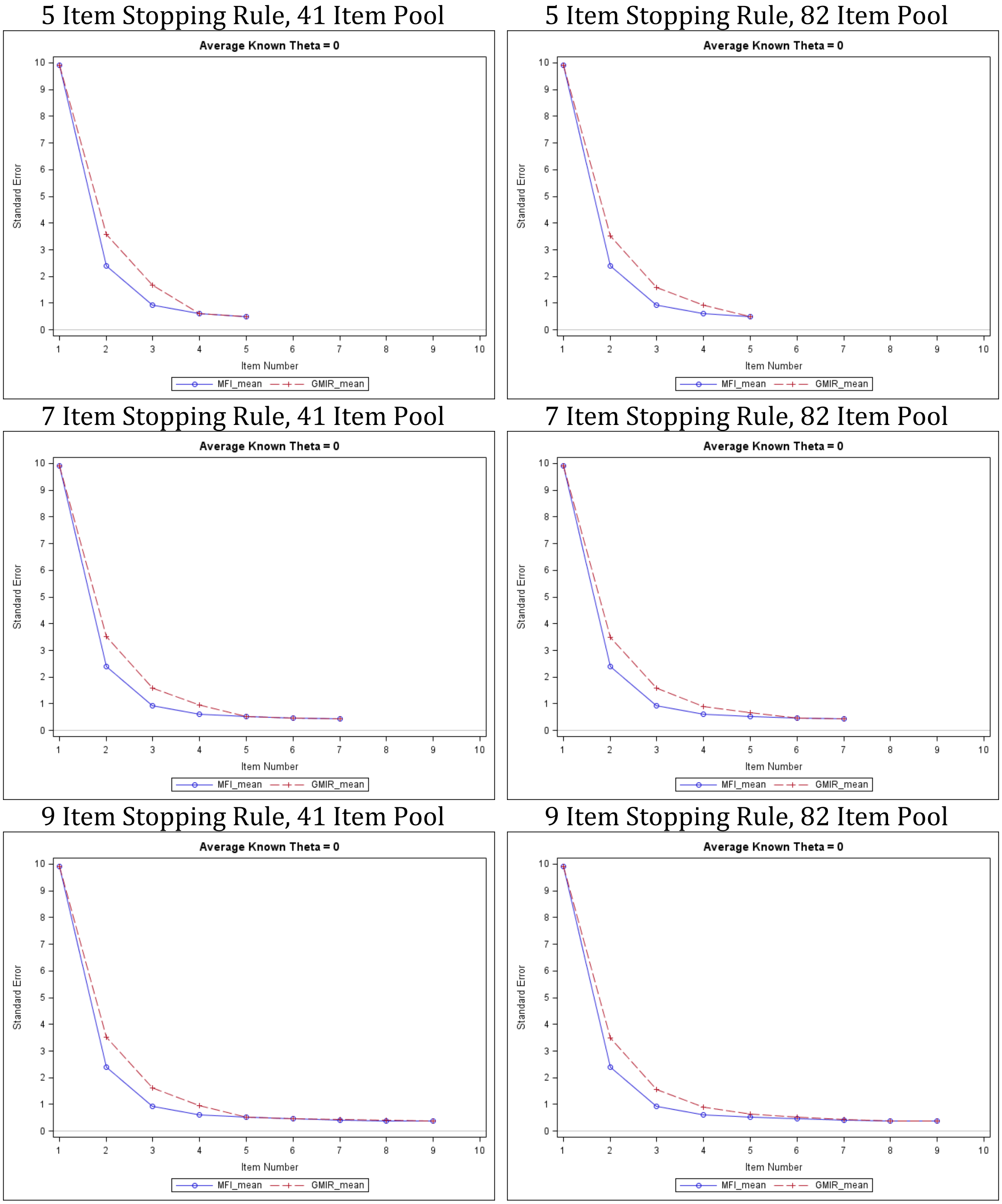
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.13. Plots of Mean SE Conditional on Item Number for Known Theta = -1 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

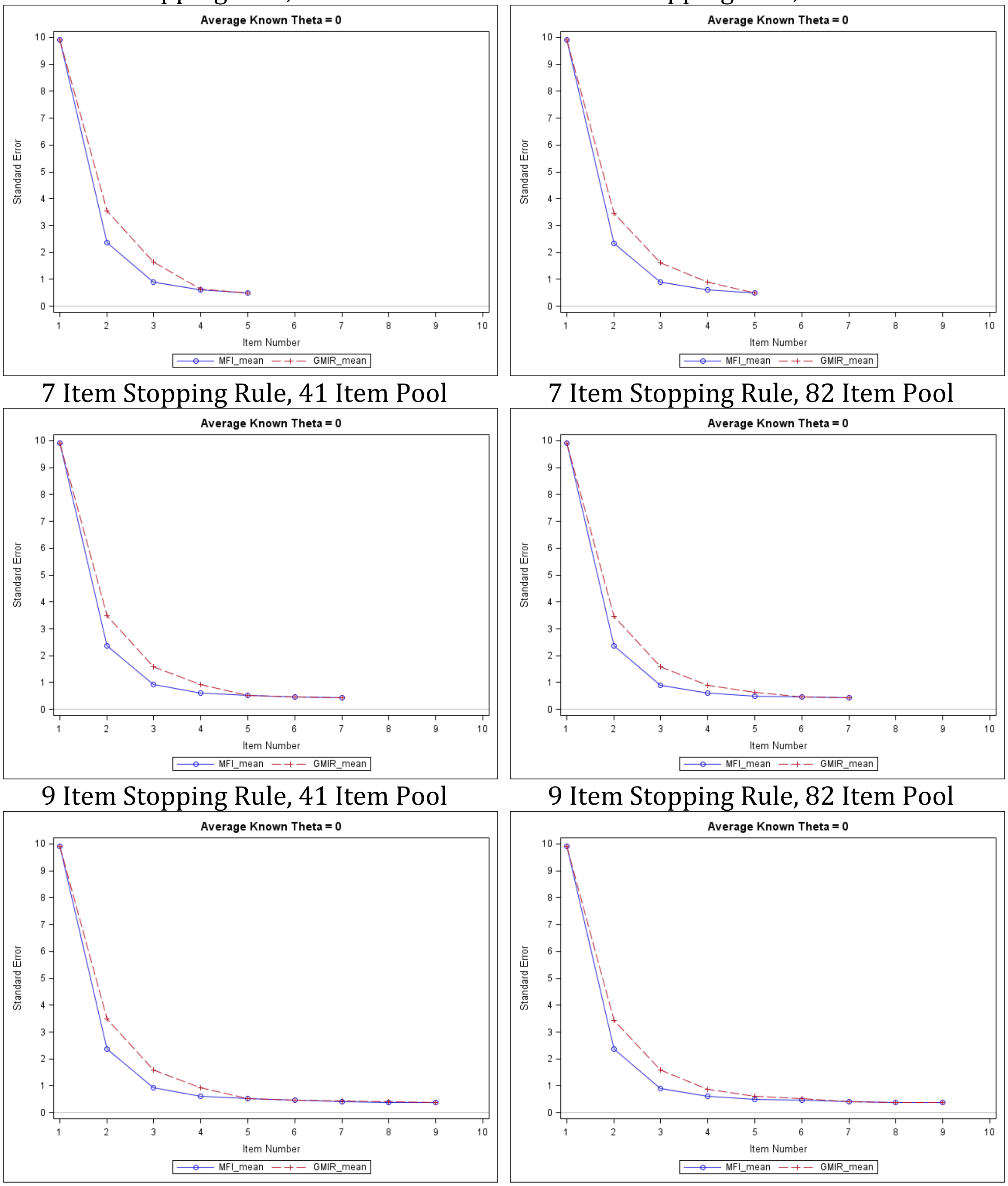
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.14. Plots of Mean SE Conditional on Item Number for Known Theta = 0 for Normally Distributed Populations



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

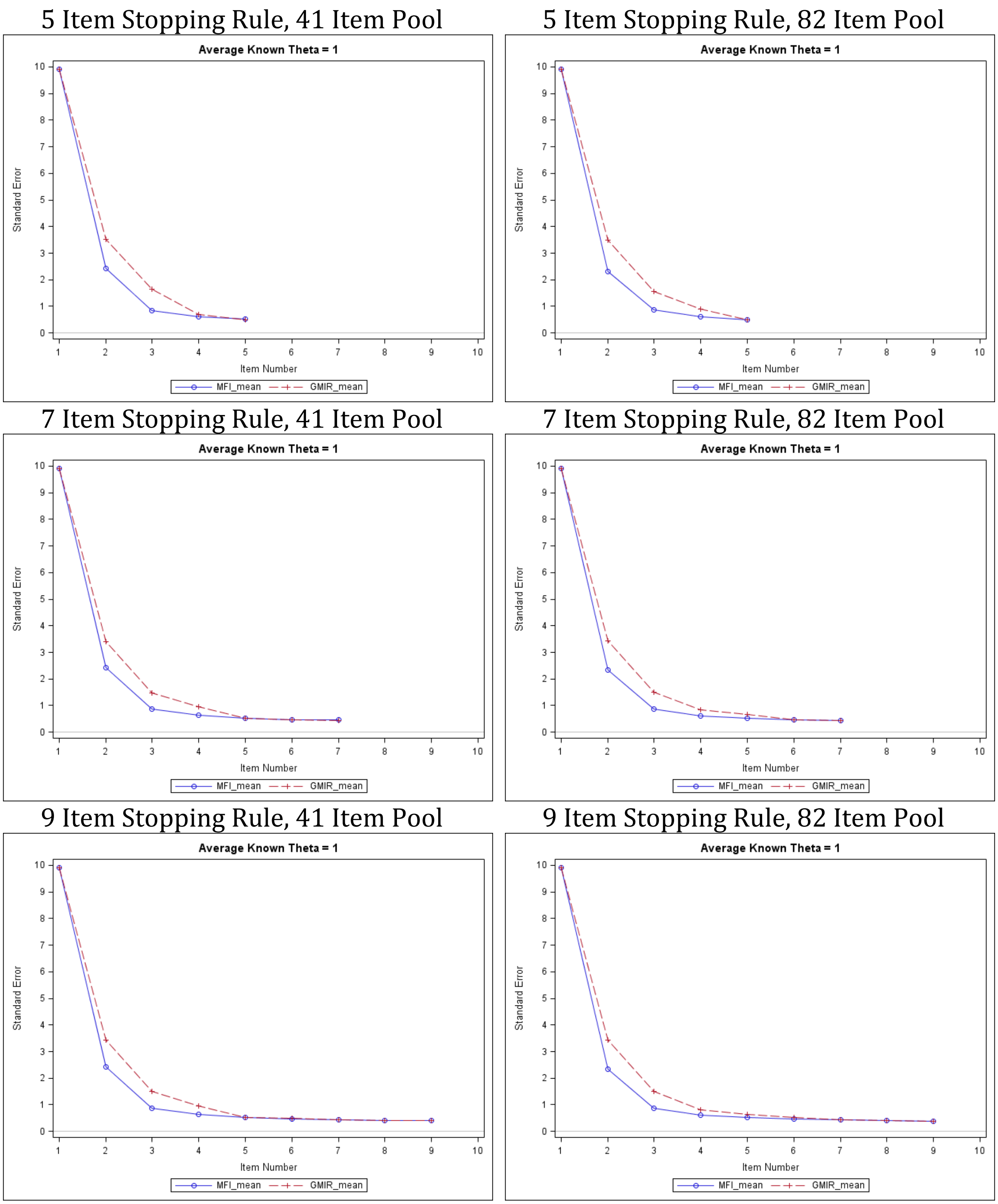
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.15. Plots of Mean SE Conditional on Item Number for Known Theta = 0 for Negatively Skewed Population Distributions



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

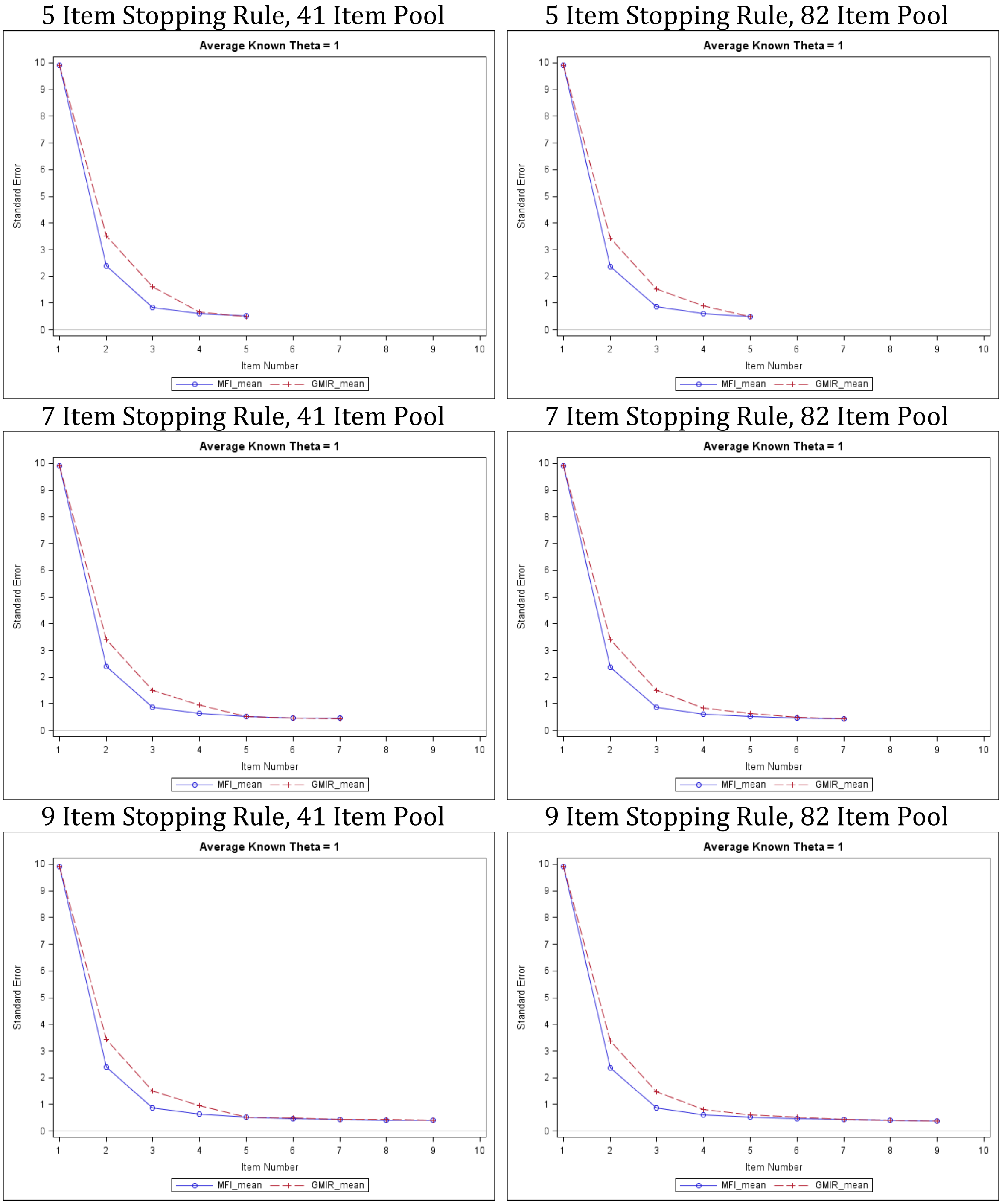
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.16. Plots of Mean SE Conditional on Item Number for Known Theta = 1 for Normally Distributed Populations


5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

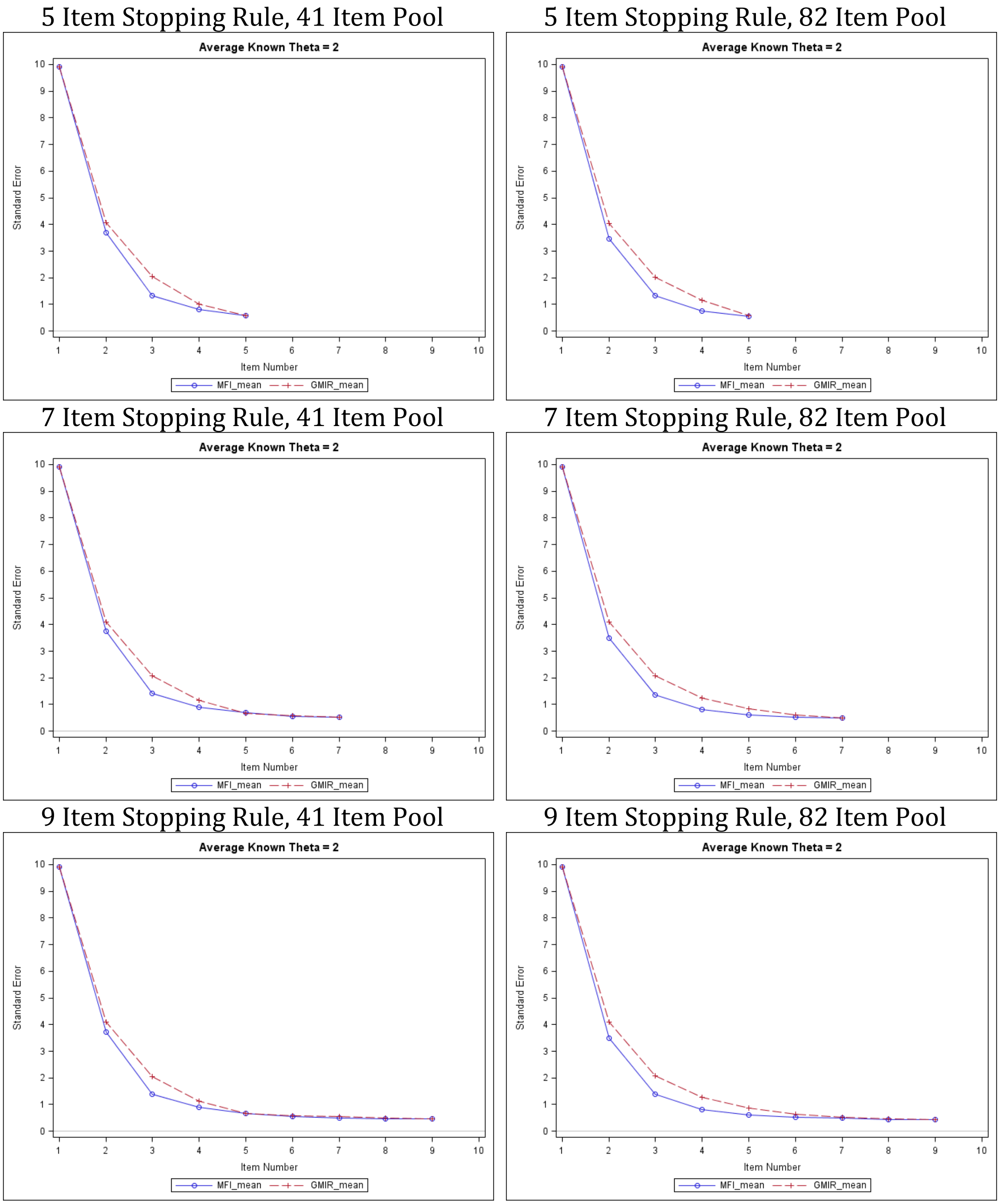
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.17. Plots of Mean SE Conditional on Item Number for Known Theta = 1 for Negatively Skewed Population Distributions


5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

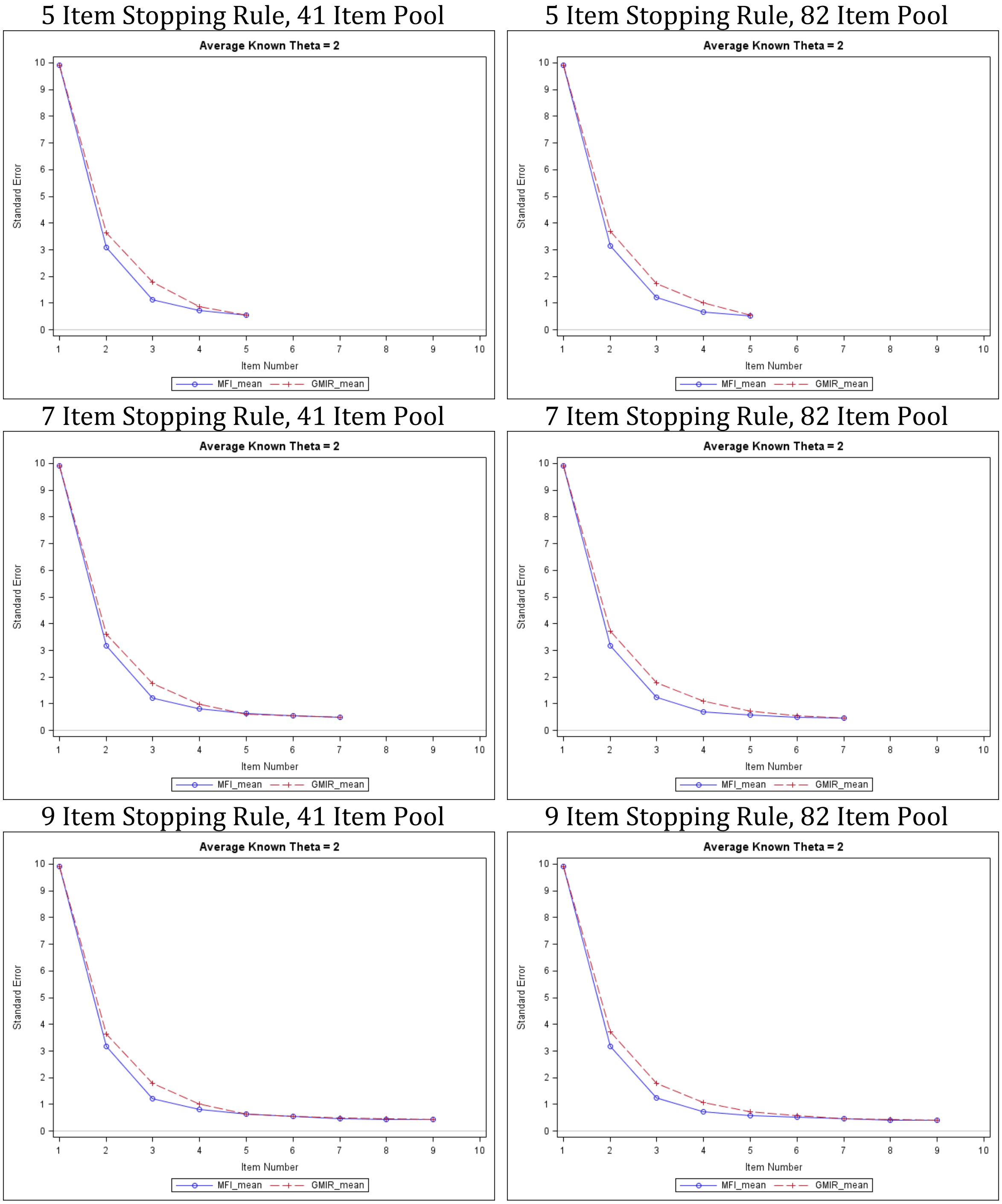
7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.18. Plots of Mean SE Conditional on Item Number for Known Theta = 2 for Normally Distributed Populations



5 Item Stopping Rule, 41 Item Pool

5 Item Stopping Rule, 82 Item Pool

7 Item Stopping Rule, 41 Item Pool

7 Item Stopping Rule, 82 Item Pool

9 Item Stopping Rule, 41 Item Pool

9 Item Stopping Rule, 82 Item Pool

Figure A.19. Plots of Mean SE Conditional on Item Number for Known Theta = 2 for Negatively Skewed Population Distributions

References

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive intervals. *Applied Psychological Measurement, 2*(4), 581-594.

Baker, B.F., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., (1972). Estimating item parameters and latent ability when responses are score in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D., (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement (pp.103-115). Hillsdate, NJ: Lawrence Erlbaum.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. In *Development and Applications of Polytomous Item Response.* Philadelphia, PA: Francis Taylor.

Burt, W., Kim, S., Davis, L., & Dodd, B. G., (2003). *Three exposure control techniques in CAT using the generalized partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229.

Chang, H.-H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222.

Chang, W. & Dodd, B. G. (2013). *A comparison of the MFI and GMIR item selection criteria in computerized adaptive testing using the generalized partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Chen, S.-K. & Cook, K. F. (2009). SIMPOLYCAT: An SAS program for conducting CAT simulation based on polytomous IRT models. *Behavior Research Methods, 41*(2), 499-506.

Chen, S.-K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement, 57*, 422-439.

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 241-255.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419-440.

Cook, K. F., Choi, S. W., Crane, P. K., Deyo, R. A., Johnson, K. L., & Amtmann, D. (2008). Letting the CAT out of the bag: comparing computer adaptive tests and an 11-item short form of the Roland-Morris Disability Questionnaire. *Spine*, *33*(12), 1378-1383.

Cook, K. F., Gartsman, G. M., Roddey, T. S., & Olson, S. L. (2001). The Measurement Level and Trait-Specific Reliability of 4 Scales of Shoulder Functioning: An Empiric Investigation. *Archives of Physical Medicine and Rehabilitation, 82*, 1558-1565.

Cook, K. F., Roddey, T. S., Gartsman, G. M., & Olson, S. L. (2003). Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. *Med Care, 41,* 823-835.

Cook K. F., Teal C. R., Bjorner J. B., Cella D., Chang C. H., Crane P. K., Gibbons L. E., Hays R. D., McHorney C. A., Ocepek-Welikson K., Raczek A. E., Teresi J. A., & Reeve B. B. (2007). IRT health outcomes data analysis project: an overview and summary. *Quality of Life Research*, *16*, 121–132.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, & J. J. Fremer (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement, 28*(3), 165-185.

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

Davis, L. L., Pastor, D., Dodd, B. G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, *4*(1), 24-

42.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement*, *49*, 789-805.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, *16*, 327-343.

Dodd, B. G. (1990). The effect of item selection procedure and step size on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Dodd, B. G., Cook, K. F. & Godin, D. G. (2005, April). *Computer adaptive medical outcome assessment: A comparison of the rating scale and successive interval models*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Dodd, B. G., & de Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol 2, pp.301-317). Norwood, NJ: Ablex.

Dodd, B. G., de Ayala, R. J., & Koch, W. (1995). Computer adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Dodd, B. G., Koch, W. R., & de Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd, B. G., Koch, W. R., & de Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*, 61-77.

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

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality Of Life Research*, *14*, 2277-2291.

Fries, J. F. (2006). The promise of the future, updated: Better outcome tools, greater relevance, more efficient study, lower research costs. *Future Rheumatology, 1*, 415-421.

Fries, J.F., Witter, J., Rose, M., Cella, D., Khanna, D., Morgan-DeWitt, E. (2014). Item Response Theory, Computerized Adaptive Testing, and PROMIS: Assessent of Physical Function. *The Journal of Rheumatology, 41*, 153-158.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*, 109-127.

Gardner, W., Shear, K., Kelleher, K. L., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, *4*: 13.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433-456.

Han, K. T. (2009, June). *A gradual maximum information ratio approach to item selection in computerized adaptive testing.* McLean, VA: Graduate Management Admission Council.

Han, K. T. (2010, May). *Comparison of non-Fisher information item selection criteria in fixed- length CAT*, Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Hinkle, D.E., Wiersma, W. & Jurs, S.G. (2003). *Applied Statistics for Behavior Sciences,Fifth Edition*. Boston, MA: Houghton Mifflin.

Ho, T., & Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed- format adaptive test. *Applied Measurement in Education, 25*(4), 305-326.

Ho, T.-H. (2010). *A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the generalized partial credit model*, (Doctoral dissertation, The University of Texas at Austin, Austin, TX).

Jette, A.M., and Haley, S.M. (2005). Contemporary Measurement Techniques for Rehabilitation Outcomes Assessment. *Journal of Rehabilitation Medicine, 27*,339-345.

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

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, *2*(4), 335-357.

Koch, W. R., Dodd, B. G., & Fitzpatrick, S. J. (1990). Computerized adaptive measurement of attitudes. *Measurement andEvaluation in Counseling and Development, 23,* 20-30.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement, 72*(1), 159-175.

LeRoux, A. J. & Dodd, B. G. (2014, April). A comparision of exposure control procedures in CATs using the GPC model. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

LeRoux, A. J., Lopez, M., Hembry, I., & Dodd, B.G. (2013). A Comparison of Exposure Control Procedures in CATs Using the 3PL Model. *Educational and Psychological Measurement, 73*(5), 857-874.

Leung, C. K., Chang, H. H. & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376-392.

Lima Passos, V., Berger, M. P., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement, 31*(3), 213- 232.

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006, April). *A variant of the progressive- restricted item exposure control procedure in computerized adaptive testing systems based on the 3PL and partial credit models*, Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*(1), 57-75.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1-20.

PROMIS Full Domain Framework. (2016, January 5). Retrieved from http://www.nihpromis.org/measures/full_framework.aspx

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practices*, *8*(3), 11-15.

Reckase, M. D. (2009). Multidimensional item response theory. New York, NY: Springer.

Reeve, B. B. (2006). Special Issues for Building Computerized-Adaptive Tests for Measuring Patient-Reported Outcomes: The National Institute of Health's Investment in New Technology. *Medical Care*, *44*(11), S198-204.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311- 327.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on traditional scaling concept. *Applied Psychological Measurement, 12*, 397-409.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph 17.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*(3), 299-311.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*(1), 57-75.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of Military Testing Association, San Diego, CA.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*(2), 201-216.

Van Rijn, P. W., Eggen, T. J., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*(4), 393-411.

Veerkamp, W. J., & Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 207-214). Tokyo: Springer-Verlag.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Routledge.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Walker, J., Böhnke, J.R., Cerny, T., Strasser, F. (2010) Development of symptom assessment utilizing item response theory and computer-adaptive testing—A practical method based on a systematic review. *Critical Reviews in Oncology/Hematology, (73),* 47-67.

Walter, O.B., Becker, J., Bjorner, J.B., Fliege, H., Klapp, B.E., & Rose, M. (2007). Development and evaluation of a computer adaptive test for "anxiety" (Anxiety-CAT). *Quality of Life Research, 16(Suppl.1),* 143-155.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, *23*(3), 263-278.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109- 135.

Ware, J. E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing. *Medical Care*, *38*, II73-II82.

Ware, J. E., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology, 50*, 71-78.

Ware, J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Daholf, C. G. H., . . . Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, *12*, 935-952

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 30*, 299-300.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.