SIMPSON, MARY ANN, Ph.D. Use of a Variable Compensation Item Response Model to Assess the Effect of Working-Memory Load On Noncompensatory Processing in an Inductive Reasoning Task. (2005)
Directed by Dr. Terry Ackerman. 203 pp.

A study of the relationship between noncompensatory processing and the working memory load of matrix completion items was conducted. Data were taken from the British Cohort Study of 1970, First Follow-up (N=14,875). To assess compensation, the GMIRT Rasch model (Spray & Ackerman, 1986), variable compensation model, was used with MCMC estimation via WINBUGS. In support of these analyses, a simulation study assessing parameter recovery for the GMIRT model was conducted. Sample size, item pool size, and interability correlation were manipulated. Adequate parameter recovery was observed when difficulty parameters were constrained equal across dimensions. In the application study, there was some evidence to support the relationship between working memory load and compensation

USE OF A VARIABLE COMPENSATION ITEM RESPONSE MODEL TO ASSESS THE EFFECT OF WORKING-MEMORY LOAD ON NONCOMPENSATORY PROCESSING IN AN INDUCTIVE REASONING TASK

by

Mary Ann Simpson

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment of the Requirements for the Degree
Doctor of Philsophy

Greensboro

2005

Approved by

____

Committee Chair

To Janet Yarbrough

And great and numerous as are the blessings of friendship, this certainly is the sovereign one, that it gives us bright hopes for the future and forbids weakness and despair. In the face of a true friend a man sees as it were a second self. So that where his friend is he is; if his friend be rich, he is not poor; though he be weak, his friend's strength is his.

- Cicero, *De Amicitia* (sec.7, Trans. E.S. Shuckburgh)

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ____

Committee Members ____

____

____

____

Date of Acceptance by Committee

____

Date of Final Oral Examination

ACKNOWLEDGMENTS

The debt owed one's graduate advisor is in the category ”debts one can never repay.” I thank Terry Ackerman for mentoring me to a successful navigation of the many hurdles in graduate study, as well as for sharing his technical expertise in psychometrics and statistical computing. I am most grateful for his listening ear when things were not going quite right. I thank Linda Wightman for inviting me to the Educational Research Department and for all of her encouragement over the years. I thank Ric Luecht for serving on my committee. He has always been most generous with his time and ideas. I thank Dan Bolt for serving and for sharing his considerable skill in multidimensional IRT and Bayesian estimation. I thank John Dunlosky for providing so much input on the cognitive theory underlying my dissertation research. I am grateful to all of my other teachers in measurement, especially Dave Thissen, Jim Laughlin, John Hattie and Lloyd Bond. I was most fortunate to have teachers of such stature in my earliest instruction in psychometrics. I thank Jeff Gill and Ryan Baker for providing such a fine course in Bayesian methods at the ICPSR Summer Institute.

I thank the staff of the Centre for Longitudinal Studies at the University of London and of the UK Data Archive for their help in using the data from the British Cohort Study. I also thank these two organizations for providing the data gratis.

I thank two graduates of our program, Sandra Neustel and Vinaya Kelkar. Both have been very supportive of me and my studies. I especially thank Sandra for nagging me to finish my dissertation and Vinaya for teaching me enough calculus to survive the impact of a course in mathematical statistic. I can't thank Anita Hawkins Dorfler enough for the yeoman service she gives to her job as department secretary. All of us in ERM owe her so much. I also thank the all of the current graduate students in Educational Research Methodology for their warm support.

I am quite grateful to my friend Janet Yarbrough for editing my dissertation and to her, Toni LaMazza and Mendel Smoak for helping me rehearse my presentation.

Graduate study can bring trying financial times with it. I am grateful to Janet Yarbrough and Fred Masheter for financial assistance, to Madison and Ruth Yarbrough for feeding me on countless occasions, and to Todd and Tammy McElroy for their gift of a 1987 Chevy Sprint when my Ford Bronco left its drive shaft in the middle lane of Interstate 85.

I wish also to thank the Rev. Lisa Fishbeck and the congregation of the Episcopal Church of the Advocate for their encouragement and prayers. Finally, I thank my two feline research assistants, Roxie and Hadley. Their insight into psychometrics may have been scant, but their coverage of the literature, particularly of whatever I was trying to read at any given time, was extensive.

PREFACE

Secondary data for this project came from the Ten-year Follow-up of the 1970 British Cohort Study. The principal investigators were J. M. Bynner and N. Butler of the University of Bristol. The funding agencies for the collection of the data were the Joseph Rowntree Memorial Trust, the Department of Education and Science, Department of Health and Social Security, Manpower Services Commission, and the National Institute of Child Health and Development(US). Data are housed in the UK Data archive at the University of Essex. Copyright is held by the Centre for Longitudinal Studies at the University of London. The original data creators, copyright holders, and the funders bear no responsibility for any analyses or conclusions presented in this document.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER I

INTRODUCTION

Background and Rationale

As problem solving human beings, we are fortunate to be able to benefit from our experience. For instance, it is one of the oldest findings in psychology that material, previously learned but unstudied for some time, is more rapidly learned on restudy than new material on first study (Ebbinghaus, 1888/2003). Also, our experience seems to be able to shield us from deficits in particular skills. For instance, experienced typists can compensate for the age related decrements in mental processing by anticipating upcoming letters (Salthouse, 1991). Such surmounting of a deficit is the generally received meaning of the term *compensation* (Baeckman & Dixon, 1992).

Tasks are probably not identical in the degree to which they allow compensation. Some tasks, such as the typing example already mentioned, may allow a great deal of compensation of one skill for another. On the other hand, other tasks may involve novel stimuli that force problem solvers to rely on a particular ability rather than on the benefits of training.

A brief *thought experiment* illustrates this idea. Figure 1 shows an eighth grade mathematics problem. The item appears to require a mixture of problem solving ability and algebra skills. However, a student, well drilled in algebra but not skillful at generating solutions in novel situations, may recognize the problem as a *rate-time-distance* problem and solve the problem by rote. Another student, with little experience in algebra, but with superior problem solving skills, may induce that speed increases as time to travel a fixed distance decreases and generate a correct answer. In this item, problem solving and algebra skills likely share a compensatory relationship. A student can compensate for lack of problem solving ability with experience in algebra, and *vice versa*. Figure 2 shows another eighth grade mathematics problem (National Center for Education Statistics, 2001). This problem appears to involve both *computational skill* and *problem solving ability*. However, the problem appears novel, not one of the typical varieties found in the math books. A student with excellent computational skills and poor problem solving skills is unlikely to answer the question correctly. No amount of memorized multiplication tables or facility in long division will rescue the situation. A student with poor computational skills but good problem solving skills is also unlikely to solve the problem. The problem requires a modest level of skill in arithmetic. In this item, problem solving ability and computational skills likely share a noncompensatory relationship. Students need to be proficient in all of the item's component skills in order to successfully answer it.

A psychometrician naturally wishes to measure these potential differences in compensation. Although the idea of measuring compensatory processes was introduced decades ago (Coombs 1964; Coombs & Kao 1955; Johnson 1935), it is only recently that estimation techniques for IRT models assessing compensation have become available. In the past five years, Bayesian MAP methods, Markov Monte-Carlo Chain (MCMC) techniques, as well as adaptations of the EM algorithm have been developed and tested on compensatory, noncompensatory, and variable compensation models (Ackerman & Turner, 2003; Bolt & Lall, 2003; Meulders, DeBoeck, Van Mechelen, 2003; Zhang, 2000, 2001). (A description of these models appears in Chapter 2.)

MCMC is a set of computational algorithms that sample from the Bayesian posterior distributions of a model's parameters. Such simulation relieves the user from calculating the integrals required by Bayesian estimation (Gilks, Richardson, & Spiegelhalter, 1996) and from calculating the derivatives required by maximum likelihood estimation. In many cases, MCMC provides the only means of estimating a model's parameters, because the required integrals and/or derivatives may not have closed form solutions.

Simulation studies are part and parcel of discovering whether, and under what conditions, a model and its associated estimation routines can survive in the world of data. Of the IRT models examined in simulation studies to date, variable compensation models have received the least attention. Variable compensation models permit assessment of the degree of compensation between the component abilities of a task. Purely compensatory or noncompensatory models do not provide for this middle ground.

A particular variable compensation model, the Generalized Multidimensional Item Response Theory model (GMIRT; Spray & Ackerman, 1986), was the focus of the present investigation. The GMIRT model includes a continuous item-level parameter, ranging from 0 (pure compensatory model) to 1 (pure noncompensatory model). Research to date has been largely exploratory. A relatively small number of items and examinees were simulated, and the compensation parameter needed to be constrained equal across items in order for any parameters of the model to be estimated (Ackerman & Turner, 2003).

The first goal of this study was to assess the effects of certain data conditions on the quality of estimation of the parameters of the GMIRT model. For practical purposes, data conditions that matter most are those affecting parameter recovery or expenditure of resources. In past simulation research on multidimensional item response models, larger sample sizes, more items, decreased correlation between abilities, and fewer dimensions were all associated with improved parameter recovery (Ackerman, Kelkar, Neustel, & Simpson, 2001a; Béguin & Glas, 2001; Bolt & Lall, 2003). In keeping with these findings, the focus of the simulation component of this study was on the effects of sample size, item pool size, and interdimensional correlation on parameter recovery. [1] The potential benefits that informative

[1] Assessing the effect of number of dimensions is a worthy goal but for practical reasons was not included here.

priors and certain parameter constraints might bring to estimation of the GMIRT model were also investigated.

Given the recency of these estimation techniques for noncompensatory models and the GMIRT model, only a handful of studies examining noncompensatory processing in tasks of interest to educational assessment have been conducted (Van Leeuwe & Roskam, 1991; Mislevy, Senturk, Almond, DiBello, Jenkins, Steinberg, & Yan, 2001). Clearly, more research on a wide variety of tasks is needed before strong conclusions about noncompensatory processes in educational tasks can be made. A second goal of the current investigation was to widen the scope of tasks that have been assessed for their degree of compensation. Previous research on compensatory processing in tasks of interest to educational assessment has focused on the LSAT (subject matter domain of questions unidentified by the authors; Van Leeuwe & Roskam) and on high school biology (Mislevy et al.).

For the current study, the task selected is more typical of the demands of intelligence testing. Specifically, a theory based assessment of noncompensatory processing in a matrix completion task was conducted. Traditionally, inductive reasoning tasks have included verbal and spatial analogies, letter and number series problems, and matrix completion problems (Pellegrino, 1985). Many view inductive reasoning as the core of what we consider *thinking* (Carpenter, Just & Shell, 1990; Garlick, 2002; Marhsalek, Lohman, & Snow, 1983; Pellegrino; Snow, Kyllonen, & Marshalek, 1984). Additionally, inductive reasoning tasks are thought to minimize the role of experience and training (Carpenter et al.). A matrix completion task was selected for the current investigation, because these tasks have an extensive research base in cognitive psychology and have, in recent years, received much attention from psychometricians (Carpenter et al.; Embretson, 2002; Hornke, 1986, 2002; Hunt, 1974; Jacobs & Vanderwerter, 1972; Ward & Fitzpatrick, 1973). Figure 3 shows an example of a matrix completion task.

If a researcher were to discover varied compensation in a set of tasks, his or her work would be only half finished. The researcher should next attempt to explain *why* these differences exist. The third goal of this study is to assess the relationship of task demands, in particular that of working memory load and compensatory processing. Past research suggests that higher cognitive loads in a task is sometimes associated with higher levels of noncompensatory processing (Billings & Marcus, 1983; Conway & Giannopoulos, 1993; Einhorn, 1971; Oglivie & Schmitt, 1979; Olshavsky, R. W., 1979; Payne, 1976). Cognitive theory suggests that the more complicated a task, the less attentional capacity remains to solve other parts of a task, to organize information, or to plan the solution of the problem (Baddeley & Logie, 1999; Engle, Kane, & Tuholski, S. W., 1999; Pellegrino & Glaser). Furthermore, for some tasks and some students, performance may be nearly *automatic*, that is, requiring little to no attentional capacity. For some other tasks or students, performance may be *effortful*, requiring much capacity. (Ackerman, P.L., 1988; Hasher & Zacks, 1979; Schneider & Shiffrin, 1977). Effortful processes are hindered by an increased working memory load; automatic processes are not (Hasher & Zacks, 1979). Because matrix problems are not heavily practiced (Carpenter et al., 1990), they are likely effortful.

A simple example can elucidate predictions concerning the relationship between working memory and compensatory processing. The matrix in Figure 3 is complicated. The examinee must discern the patterns governing two sets of attributes — shading and shape names. He or she must also keep mental track of his or her own plans for and progress in solving the problem. In other words, this matrix has a high working memory load. A student not adept at keeping track of his or her own problem solving is unlikely to compensate with a remarkable ability to discern patterns. In short, we, therefore, expect an increase in noncompensatory processing for items with an increased working memory load.

Overview of Method

A simulation study was conducted to evaluate the effects of sample size, interability correlation and other constraints on parameter recovery for the GMIRT model. The analyses of the matrix completion items, the application study, relied on cognitive theory. Over the past 25 years, psychometricians have made progressively greater use of ideas from cognitive psychology (Carroll, 1976; Pellegrino & Glaser, 1979; Mislevy, 1995; Newell & Simon, 1972; Sternberg, 1977; Whitely & Schneider, 1980). Today's researchers have numerous exemplars for cognitive task analyses.

In particular, Carpenter et al.'s (1990) cognitive task analysis of the *Ravens Advanced Progressive Matrices* provided the model for a cognitive task analysis of the matrices subscale of the British Ability Scales (BAS; Elliott, 1978), the English cousin of the better known Differential Ability Scales (Elliott, 1990). The cognitive task analysis of the BAS matrices were then expressed mathematically as an instance of the GMIRT model. MCMC estimation of model parameters was conducted. Statistics were calculated on overall model fit. Descriptive comparisons of how the GMIRT, noncompensatory, and compensatory models differ concerning person-fit were also conducted. Full details of the method appear in Chapter 3.

Importance of Study

First, this study should advance our understanding of the cognitive processes involved in a particular inductive reasoning task. Basic research is not a ”dirty word.” Second, this study is exceptionally timely. The influx of cognitive psychology into psychometrics has reached large scale testing, as shown in recent works by DiBello (2002) and Hartz, Roussos, and Stout (2002), and by the interest general testing is taking in the Evidence Centered Design (ECD) approach to test construction (Mislevy, Steinberg, & Almond, in press). Third, the findings may be of practical use to test developers. Test developers may wish to create items in which compensation for a skill deficit is not likely to lead to success on the item (e.g., relying on superior vocabulary to compensate for problems in reading comprehension). Fourth, if compensatory processing is, indeed, hindered by increased working memory load, this finding may help curriculum planners to design materials appropriate for students with learning disabilities. Planners may try to lighten the working memory load of tasks, so that learning disabled students with low working memory capacity can actually apply compensatory strategies that their teachers taught them. Fifth, the study should increase understanding about what works and does not work in MCMC estimation of complex IRT models. Finally, the study will present another demonstration that sophisticated measurement models can be of use to cognitive psychologists.

Summary of Research Questions

To recapitulate, the three main research questions of this study were:

1. What effect do sample size, item pool size, interdimensional correlation, and various parameter constraints have on the quality of estimation of the compensation parameter in the GMIRT model?

2. To what degree does noncompensatory processing operate in an inductive reasoning task?

    a. Is the cognitive processing required by some items better explained by a model permitting less compensation than by a pure compensatory model?

    b. For individuals whose performance is poorly explained by a compensatory model, is their performance better explained by a model allowing less compensation?

3. To what degree does noncompensatory processing change as the working memory load of an item increases? Given that coordinating the skills involved in the matrix reasoning task is likely an effortful process, we expect an increase in noncompensatory processing as the working memory load increases.

CHAPTER II

REVIEW OF THE LITERATURE

Both measurement theory and cognitive theory motivate the current study. Background information on both is provided below. The topics in measurement theory to be covered are: logical, mathematical, and empirical foundations of measuring compensatory and noncompensatory processes; Bayesian statistics in general; and MCMC estimation in IRT. Topics in cognitive theory to be covered are: cognitive theory for inductive reasoning tasks and working memory. Previous research on compensatory and noncompensatory processes as well as research on the relationship of noncompensatory processes and working memory will also be discussed.

Logical and Mathematical Foundations of Measuring Compensatory and Noncompensatory Processes

Because prepositional logic provides much of the terminology for noncompensatory models, a brief explanation of its major tenets is provided here. Noncompensatory models go by many other names, chief of which is *conjunctive*. *Conjunction* refers to the joining of two or more propositions by the word *and*, for instance, p $\wedge$ q, or ”John is a scholar *and* John is a gentleman”. The term *disjunction* refers to the joining of two or more propositions by the word, *or*, for instance, p $\vee$ q, ”John is either a scholar *or* John is a gentleman”. Conjunctions are true if and only if all component propositions are true. Disjunctions are true if and only if any of the component propositions are true (Copi, 1953; Whitesitt,1961).

These concepts have been eagerly applied to problem solving, for correct solution of a problem is conceptually equivalent to a true proposition. Coombs (1955,1964) presents several models of problem solving including the conjunctive, the disjunctive, and the compensatory models. In Coombs' theory of scaling, both persons and tasks can be ordered particular ability dimensions. In a conjunctive task, the person must surpass the tasks' values on all of the ability dimensions (e.g., math problem in Figure 2). In a disjunctive task, a person need only surpass the task's value on at least one of the ability dimensions.

Coombs and Kao (1955) defined three compensatory models, the *stimulus* compensatory model, the *individual* compensatory model, and the *jointly* compensatory model. The stimulus compensatory model involves persons' responding to a fixed stimulus. The task determines the level of performance necessary for success. To succeed on the task, an individual must possess some weighted combination of the required abilities exceeding the value of the stimulus' weighted combination of these abilities (Coombs, 1964; Coombs & Kao, 1955). Figure 4 (adapted from Coombs, 1964) plots the ability requirements of a task against the ability profiles of two examinees. Scores on ability 1 are plotted on the x-axis, and scores for ability 2 are plotted on the y-axis. The vector q_1 represents the ability demands of a hypothetical task. Vectors p_1 and p_2 plot two examinees' ability profiles. The length of the vector reflects the total amount of ability possessed by an examinee or required by the item. Examinee p_1 falls short, not having enough of the ability composite to pass the item, whereas examinee p_2 has more than enough of the ability composite to pass the item. (Chapter III presents mathematical details on constructing vectors representing task demands; Figure 4 is mainly an heuristic device). The stimulus compensatory model's best known manifestations are multiple regression analysis and item factor analysis (Coombs & Kao, 1955). Additionally, most of the IRT models in current use assume a stimulus compensatory model.

The individual compensatory model involves a fixed individual responding to stimuli (Coombs,1964; Coombs & Kao, 1955). The person determines the threshold for the task's success. Coombs gives the example of a camper selecting travel items based on their utility and lightness. For instance, different individuals may assign different values to the utility of bringing a CD player versus that of taking a weather radio. Decision-making models are a prime example of the individual compensatory model. Individuals rate stimuli according to the importance of particular features, e.g., price, durability, attractiveness. For instance, Figure 5 (adapted from Coombs, 1964) shows vector p_1 representing a person valuing two stimulus attributes equally and vectors q_1 and q_2 representing stimuli possessing different combinations of these attributes. Stimulus 2 fails the person's threshold, whereas stimulus 1 passes it.

In the *jointly* compensatory model, both the stimuli and the persons have a share in determining what combination of abilities constitutes success (Coombs, 1995). Coombs does not develop this model and appears to have omitted it entirely in his 1964 monograph. The concept of a jointly compensatory model does, however, point to a tension in psychometrics. Researchers often must decide whether an attribute is a feature of the item, the examinee, or of both items and examinees.

Latent Variable Probabilistic Models

Coombs (1955) emphasizes that his models are completely deterministic. In other words, items and persons function perfectly according to the underlying model (Gregory, 1992). Figure 6 shows a response surface for a deterministically compensatory item with two equally weighted dimensions. Again, the model involved is the stimulus compensatory model. In this particular case, the item difficulty has been scaled to 0. It is clear from the graph that any combination of ability scores averaging more than the item difficulty of 0 will result in a correct answer. On the other hand, Figure 7 shows a contour plot of a response surface for a deterministically noncompensatory item with two equally weighted abilities. Here, it is clear that both ability scores must be greater than the 0 threshold for the item difficulty for a correct answer to result.

Numerous probabilistic models for compensatory and noncompensatory processing have been developed over the past few decades. In a particularly insightful review, Junker

(1999) divides all of these latent variable models into two groups, those modelling abilities as continuous variables and those modelling abilities as binary matrices. Junker holds that the difference in usage between them is one of granularity. The continuous models are usually employed when a broad description of a few dimensions is needed. The binary matrix models are usually employed when a description of numerous components is needed. The distinction between these groups of models has blurred somewhat in recent years.

Continuous Variable Models The simplest compensatory and noncompensatory models will be discussed first. Equation 1 shows a multidimensional Rasch model (adapted from Embretson & Reise, 2000; McKinley & Reckase, 1982),

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, d_i) = \frac{exp(\sum_{k=1}^{m} \theta_{pk} + d_i)}{1 + exp(\sum_{k=1}^{m} \theta_{pk} + d_i)}, \tag{1}$$

where Y is the observed 0, 1 response to item i, $\boldsymbol{\theta}$ the vector of the p_{th} person's abilities on the underlying k dimensions, and d_i the overall difficulty for item i. The probability of a correct response increases as the average of the component abilities also increases. Each of the abilities carries equal weight in the Rasch formulation (Embretson & Reise). Item difficulty is a scalar quantity, conventionally labeled d_i with high positive values indicating *easier* items. The kernel $\theta_{pk} + d_i$ in the compensatory model is analogous to Lord's parameterization, $\theta_p - a_i b_i$, in unidimensional IRT models (Ackerman, Kelkar, Neustel, & Simpson, 2001b). Figure 8 shows a contour plot for a two-dimensional instantiation of this model. The contour lines show probabilities of correct response in equal intervals, usually .10 wide, as a function of ability 1 (x-axis) and ability 2 (y-axis). The compensatory nature of the model is readily apparent. A person with a score of 3 on one ability and -3 on the other still has a 50/50 chance of answering the item correctly.

Equation 2 shows the multidimensional noncompensatory Rasch model (Whitely, 1980; Maris, 1995).

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, \mathbf{b_i}) = \Pi_{k=1}^{m} \frac{exp(\theta_{pk} - b_{ik})}{1 + exp(\theta_{pk} - b_{ik})} \tag{2}$$

This model has also been called the Multicomponent Latent Trait Model (MLTM) (Embretson, 1980; 1986). In this and other noncompensatory models, the probability of a correct response increases only as both abilities increase. In contrast to the compensatory model, the difficulty parameter in the noncompensatory model, $\mathbf{b_i}$, is a vector. There is a separate difficulty parameter for each of the k components. The difficulty parameters in the noncompensatory model are usually labeled b_{ik} with high positive values indicating more *difficult* items. Additionally, the difficulty parameters in the noncompensatory models are not confounded with discrimination as in the compensatory models.

Figure 9 shows a contour plot for a two-dimensional instantiation of this model. The components' difficulty parameters have both been set to 0, and both abilities are equally weighted. Here, a person with a score of 3 on the first ability would need at least 0 on the second ability in order to have at least a .50 probability of answering the item correctly.

These basic compensatory and noncompensatory latent trait models have been expanded to include additional parameters for discrimination, $\mathbf{a_i}$, and guessing, c_i (Ackerman, 1994, 1996; McKinley, 1989; Reckase, 1985; Reckase & McKinley, 1991). Equation 3 shows the three-parameter logistic compensatory model (adapted from Ackerman, 1996).

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, d_i, \mathbf{a_i}) = c + (1 - c) \frac{exp(\sum_{k=1}^{m} a_{ik}\theta_{pk} + d_i)}{1 + exp(\sum_{k=1}^{m} a_{ik}\theta_{pk} + d_i)} \tag{3}$$

Equation 4 shows the three-parameter noncompensatory model (Sympson, 1978).

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, \mathbf{b_i}, \mathbf{a_i}, c_i) = c_i + (1 - c_i)\Pi_{k=1}^{m} \frac{exp(a_{ik}(\theta_{pk} - b_{ik}))}{1 + exp(a_{ik}(\theta_{pk} - b_{ik}))} \tag{4}$$

The discrimination parameters in both compensatory and noncompensatory models may be interpreted much like factor loadings; high positive values imply more of the construct, and low values imply less of the construct. The relationship between the discrimination parameters in the compensatory MIRT model and factor loadings is exact (Equation 5; McDonald, 1999).

$$a_{ik} = \frac{\lambda_{ik}}{\sqrt{1 - {\lambda_{ik}}^2}} \tag{5}$$

Figure 10 shows a contour plot for the two-parameter compensatory model for an easy item measuring almost entirely the first dimension. The contour lines show response probabilities in equal intervals as a function of ability 1 (x axis) and ability 2 (y axis). The contour lines run perpendicular to the direction of optimal measurement for the item (Ackerman, 1996). Figure 11 shows a contour plot for the two-parameter compensatory model for a difficult item measuring almost entirely the second dimension. Figures 12 and 13 show the noncompensatory contour plots for items with the same discrimination and difficulty parameters as in the previous compensatory items.

Several authors have proposed multidimensional compensatory models based on the cumulative normal distribution (Bock, Gibbons, & Muraki, 1988; McDonald, 1967, 1997). Equation 6 shows the two-parameter multidimensional normal ogive model (adapted from Embretson and Reise, 2000). The upper limit of integration is the threshold value for the item, and z is the standard score corresponding to the item's threshold value.

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, b_i, \mathbf{a_i}) = \int_{-\infty}^{\sum_{k=1}^{m} a_{ik}\theta_{pk}+d_i} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right) dt \tag{6}$$

Estimation of the parameters for compensatory logistic models. Models are pretty. However, without ability and item parameter estimates, they are pretty useless. Psychometricians have attempted a variety of methods to estimate the parameters of the compensatory and noncompensatory IRT models. Maximum likelihood methods have been successfully used to estimate person and item parameters for many of the compensatory logistic models described above. However, none of these models has closed-form derivatives (Seagall, 1996), so all maximum likelihood methods require the aid of iterative techniques, such as the Newton-Raphson or Fisher techniques. Bayesian expected a postieri (EAP) and maximum a postieri (MAP) estimation of the parameters of the multidimensional logistic models have enjoyed their widest application in multidimensional adaptive testing. Also, MCMC methods have also been successful with the logistic compensatory models. Bolt and Lall (2003) found adequate performance of MCMC parameter recovery for a two-dimensional compensatory model. The results held over a mix of sample sizes, item pool sizes, and interability correlations. However, Bolt found that the NOHARM (Fraser, 1988) program offered residuals consistently smaller than those of the MCMC procedure.

Eliminating multidimensionality by creating a unidimensional composite score is a surprisingly little used technique for making sense of multidimensional data. Once the composite is created, standard techniques for unidimensional IRT models can be used to calculate parameter estimates. This composite is usually mathematically determined (Ackerman, 1994; Bloxom & Vale, 1987; Luecht, 1996; Wang, 1986, 1987; Van der Linden, 1996). Small-scale simulation studies support the idea that the reference composite's parameters can be adequately recovered by standard unidimensional IRT estimation techniques (Wang, 1986, 1987).

Estimation of the parameters for the compensatory normal ogive model. Baker (1992) acknowledges that the logistic latent trait model is far simpler to estimate than the normal ogive model. However, the only software for multidimensional item response theory in wide use today, NOHARM (Fraser, 1988; Fraser & McDonald, 2003) and TESTFACT4 (Wood et al., 2003), estimates the parameters for the normal ogive model. Both programs employ nonlinear factor analysis (see Christoffersson, 1975; DeChamplain, 1999; McDonald, 1967, 1999, for in-depth discussion). Equation 7 shows the normal ogive model for nonlinear factor analysis.

$$p(Y_{pi} = 1|\boldsymbol{f_p}, \tau_i, \boldsymbol{\lambda_i}) = \int_{-\infty}^{\frac{\sum_{k=1}^{m} \lambda_{ik} f_{pk} - \tau_i}{u}} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-z^2}{2}\right) dt \tag{7}$$

The $\boldsymbol{\lambda}$ represents the matrix of factor loadings, $\mathbf{f}$ the vector of factor scores, u the uniqueness, and τ_i the response threshold for that item (McDonald, 1999). The response threshold shares an exact relationship with item difficulty (Equation 8; Fraser & McDonald, 2003; McDonald, 1999).

$$\tau_i = \frac{-d_i}{\sqrt{1+u}} \tag{8}$$

MCMC has been successfully applied to the multidimensional normal ogive model as well as to the multidimensional logistic compensatory model. Béguin and Glas (2001) found that MCMC performed well across a range of priors. In almost all cases, parameter recovery was similar to that of NOHARM (Fraser, 1988) and TESTFACT (Wilson, Wood, & Gibbons, 1998). The exceptions involved situations in which the priors were quite different from the generated parameters.

Estimation of the parameters for noncompensatory logistic models. Estimation of the noncompensatory models has a shorter, but far more tortured history. Many serious attempts failed (Ackerman, 1989; Hsu, 1995; Lim, 1993). To date, only two non-MCMC approaches appear promising. Both make use of the EM algorithm. The first approach (Maris, 1992, 1995) has been applied only to the the noncompensatory Rasch model (MLTM); the second has been applied to the two- and three-parameter noncompensatory model and would, presumably, work for the noncompensatory Rasch model (Zhang, 2000, 2001). Maris' (1992, 1995) approach models performance on the latent components as missing data in the EM algorithm (Embretson, 2000). Maris derives both ML and MAP estimates for the noncompensatory Rasch model. However, the ML methods failed to yield finite estimates. As Maris (1995) reported, there have yet to be large-scale simulation studies assessing this method's parameter recovery under various data conditions.

Zhang (2000, 2001) incorporates a genetic algorithm as the M step in EM estimation. The genetic algorithm is designed to ”mimic the principles of natural evolution” (Zhang, 2000, p. 9). In language of evolution, the *fittest* parameter estimates *survive* and *breed*. In the language of statistics, parameter estimates resulting in larger values for the posterior expectation function are deemed the *fittest*. They are retained into the next iteration, i.e., *survive*. Some estimates then have random quantities added to their values, i.e., *mutate*. Surviving estimates then exchange their values with those of other population members , i.e., *breed*. The process repeats until particular convergence criteria are met.

Parameter recovery for compensatory models in the genetic algorithm compare favorably with that of NOHARM (Zhang, 2001). Results were also favorable in the small amount of MCMC research to date concerning noncompensatory models. Bolt and Lall (2003) found that parameter recovery for the noncompensatory Rasch model was adequate when the correlation between dimensions was small (0 or .3). Increasingly larger numbers of items and persons were needed to insure adequate parameter recovery when the interability correlation was larger (.6). However, for all conditions simulated, parameter recovery of the logistic compensatory model remained superior to that of the noncompensatory Rasch model.

Variable compensation models. The noncompensatory logistic model can be rewritten as a set of main effects and interactions. Equation 9 shows this reformulation for a two-dimensional noncompensatory model (Ackerman & Bolt, 1995).

$$p(Y_{pi} = 1|\boldsymbol{\theta_p}, b_i, \mathbf{a_i}, \mu_c) = \frac{exp(f_1 + f_2)}{1 + exp(f_1 + f_2) + \mu_c(exp(f_1) + exp(f_2))} \tag{9}$$

The f_1 and f_2 terms are representations for $a_1(\theta_1 - b_1)$ and $a_2(\theta_2 - b_2)$ respectively. The μ_c term is the weight applied to the interaction of θ_1 and θ_2. When the compensation parameter is 0, the compensatory model results. When it is 1, the noncompensatory model results. The compensation parameter can take on any value between 0 and 1 to reflect varying degrees of compensation between the abilities required by a task. One might also conceptualize this compensation weight to be the probability of successfully applying a purely noncompensatory combination of abilities to a task. This model is referred to as the Generalized Multidimensional IRT (GMIRT) model (Spray & Ackerman,1986; Ackerman & Bolt). The $\boldsymbol{\mu_c}$ parameter in the original formulation of the model is fixed across persons and variable across items [2]. An additional assumption of the model is that the amount of compensation between the abilities is equal across abilities, e.g., the amount of compensation between ability 1 and 2 in a three-ability model is the same as the amount of compensation between abilities 1 and 3, and 2 and 3. The GMIRT has recently been successfully estimated and applied to data via MCMC methods (Ackerman & Turner, 2003). Figures 14, 15, and 16 show probability contour plots for an item with equal discrimination parameters, average difficulty, and μ_c set to .20, .50, .80 respectively.

Binary Matrix and Hybrid Models Binary matrix models are discussed in detail because they are nearly always conjunctive models. The core of such models is two matrices: the first, an $i\,x\,k$ binary matrix representing an examinee's possession or nonpossession of the several attributes required to solve an item; the second, an $i\,x\,k$ matrix representing each item's required skills. The first matrix is usually named α, or ξ when referring to only one item. The second matrix is usually named Q (Tatsuouka, 1995). The entries in the α matrix are 1, if an examinee possesses the attribute, or 0 otherwise. The entries of the Q-matrix are 1, if an

[2] It is entirely plausible to model compensation as a random effect across persons and items. Such an arrangement would be most helpful in estimating individual differences in compensatory processing. However, the estimation difficulty would increase dramatically. For now, compensation is modeled as fixed across persons.

attribute is required, or 0 otherwise. A response to an item will be correct, if the examinee possesses all of the attributes required by that item. It is readily apparent that this is a conjunctive model. Table 1 shows a hypothetical Q-matrix and Table 2 a hypothetical α matrix. Under this model, our examinee would answer only item 3 correctly. The researcher incorporates cognitive theory and/or professional expertise in developing the Q-matrix.

For practical use, binary matrix models models are made probabilistic. Nearly all binary matrix models can be built from the following simple model (Junker, 1999; 10).

$$p(Y_{pi} = 1 | \xi_{pi}, s_i, g_i) = \xi_{pi}(1 - s_i) + (1 - \xi_{pi})g_i \tag{10}$$

The term ξ_{pi} is set to 1 if the examinee possesses all of the skills required by item i and 0 otherwise. The s_i is a *slip* parameter. *Slippage* occurs when a person possesses the component in question, but does not apply it. The g represents use of alternative strategies such as guessing. The probability of a correct response is the sum of the probability of possessing and using the specified attributes and the probability of lacking attributes and solving the problem with an alternate strategy. IRT functions are often incorporated into pure binary matrix models, resulting in hybrids of the continuous variable and binary matrix models.

A brief discussion of rule space approaches, probability matrix decomposition (PMD) models, and hybrid models follows.

Rule space approaches. Tatsuouka (1995) identifies the main elements of rule space modeling as analysis of knowledge structures and classification of an individual's performance into predetermined knowledge states. In practice, these elements involve creation of a Q-matrix and classification of examinees into mastery/nonmastery groups or into groups distinguished by particular problem solving techniques (i.e., rules). Common rule space models include the unified model (DiBello, Stout, & Roussos, 1995) and its sibling, the fusion model (Hartz, Roussos, & Stout, 2002). Both the unified and fusion models include a potentially multidimensional residual ability parameter. This parameter is intended to describe unmodeled yet important attributes (Hartz et al; Di Bello et al.). Equation 11 shows the unified model (Hartz et al.).

$$p(Y_{pi} = 1|\boldsymbol{\alpha}_{pi}, \theta_{pi}) = d_i \Pi_{k=1}^{m} \pi_{ik}^{\alpha_{ik} q_{ik}} r_{ik}^{(1-\alpha_{ik}) q_{ik}} p_{c_i}(\theta_p) + (1 - d_i) p_{b_i}(\theta_p) \qquad (11)$$

The term d_i is the probability of the examinee choosing the strategy specified in the Q-matrix. The term π_{ik} is the probability of the examinee correctly applying a required attribute, given that the attribute has been mastered. The term r_{ik} is the probability of the examinee correctly applying a required attribute, given that the attribute has not been mastered. The term q_{ik} is the item's entry in the Q- matrix for attribute k. The term $p_{ci}(\theta_p)$ is the Rasch IRT function with the difficulty parameter, c_i. In this model, the c parameter represents the *cognitive completeness* of the Q-matrix. The more difficult it is to solve the item with the residual ability, the more *complete* the Q-matrix is considered. The term $p_{b_i}(\theta_p)$ is the Rasch IRT function with the difficulty parameter b representing the difficulty of using a different strategy for solving the item. Again, the θ_p in the model is a residual ability parameter.

The parameters of the unified model have been successfully estimated by MML and MCMC methods, albeit with numerous constraints (Jiang & DiBello, 1996; Hartz et al., 2002). The results of the MCMC analyses have been presented at public conferences but are still not yet widely available (Hartz et al.). The authors' MML method incorporates the EM algorithm with the genetic algorithm in the M step. According to the publications and conference presentations released to date, the results of the MML estimation have been mixed (DiBello, Stout & Roussos, 1995; Jiang & DiBello). The authors' final version of the estimation algorithm is due to be published shortly.

The parameters of the fusion model have been successfully estimated by MCMC methods, aided by several proprietary programs to attain pre-estimates of some of the parameters and then pass them into the MCMC routines as fixed values. The fusion model estimation routines have shown adequate MCMC convergence, reliability, robustness to minor misspecifications of the Q-matrix, and good classification of examinees into mastery/nonmastery groups for elements in the Q-matrix (Hartz et al., 2002; Henson, Stout, He, & Douglas, 2004).

Probability matrix decomposition models. The PMD models are similar to the rule space models. Like the unified and fusion models, the PMD models incorporate a latent variable for correct application of task components and another for examinees' mastery of components. Unlike the rule space approaches, the PMD models do not require a prespecified matrix delineating which items require which skills. The user can perform exploratory analyses and only need to know beforehand the number of latent attributes. Additionally, PMD researchers have experimented with response functions other than the conjunctive model (Maris, 1999; Meulders et al., 2003).

PMD models assume several parameterizations. Because of restrictive statistical assumptions, earlier PMD models, such as the Multiple Classification Latent Class Model (MCLCM; Maris, 1999), can not incorporate between-person variation in component mastery or component application (Meulders et al., 2003). In other words, these effects are fixed across persons. The more recent PMD models can incorporate such between-person variation (Meulders et al.). The earlier, constrained, PMD models have enjoyed reasonably successful estimation under EM and MAP approaches (Maris). However, the more recent models have shown only limited success in estimation via EM and MCMC (Meulders et al.).

Hybrid models. In some models, the role of the continuous variable IRT function is far more prominent than in the unified and fusion models. The factor-analytic qualities of the continuous models and componential qualities of the binary matrix models become joined. This seems less strange when one considers that the Q-matrix can be viewed as confirmatory factor loadings without the requirement of simple structure (For a similar discussion, see Junker, 1999). Adams, Wilson, and Wang's (1997) multidimensional random coefficients logit model (MRCLM) is a prime example of a hybrid model. In the simplest form of the MRCLM, the researcher creates a binary weight matrix for abilities required to solve an item and then fixes the $\mathbf{a_i}$-parameters to the values in this matrix. This weight matrix is identical conceptually to the Q-matrix. The MRCLM has been successfully estimated via the EM algorithm (Junker). An additional method of blending binary and continuous models is to treat the entire model as a second-order factor model with a single overarching continuous ability as the parent variable to all of the mastery/nonmastery indicators (de la Torre & Douglas, 2004).

Model for the Current Study As the reader can appreciate, there are a plethora of latent variable models for compensatory and noncompensatory processes. Which model or combination of models should be chosen for the current study? There are three burdens the selected model must shoulder. First, it should be able to assess the degree of compensation between latent abilities. Second, it should be able to incorporate a prespecified model of cognitive processing for the matrix completion task. Third, it must provide person- and item-parameter estimates accurate enough for research purposes.

The GMIRT model can be modified to include a design matrix (Q-matrix, fixed $\mathbf{a}$-parameters) along the lines of the MRCLM (Adams et al., 1997). The GMIRT model's discrimination parameters can be prespecified 0 or 1 as in the Q-matrix, and the degree of compensation between component abilities can be assessed. Whether this hybrid model meets the third requirement, that of estimation accuracy, is, of course, an empirical question, the answer to which is one of the major goals of the current study.

Bayesian Statistics and IRT

Because MCMC techniques will be used to estimate the parameters of the GMIRT model, a detailed discussion of Bayesian statistics in general and its use for IRT models in particular is given here.

Bayesian statistics attempts to formalize and quantify researchers' prior assumptions concerning their research questions. The *answers* in Bayesian statistics, the posterior distributions, represent the relationship between observed data and the prior assumptions (Gill, 2002). To quantify prior assumptions, a researcher places a probability distribution on them. To quantify purported knowledge concerning the observed data, a researcher describes them with a traditional likelihood function. Mathematically, the posterior distribution for a parameter is given by Equation 12. The $p(\mathbf{y})$ is a constant that insures the quantity is a bona fide probability. In most situations, it can be omitted, and the posterior distribution expressed as in Equation 13 (adapted from Gelman, Carlin, Stern, & Rubin, 1995). In this discussion, π denotes *posterior probability*, and *p* probability in general or *prior* probability.

$$\pi(\mu|\mathbf{y}) = \frac{p(\mu)p(\mathbf{y}|\mu)}{p(\mathbf{y})} \tag{12}$$

$$\pi(\mu|\mathbf{y}) \propto p(\mu)p(\mathbf{y}|\mu) \tag{13}$$

For some problems, it is reasonably simple to calculate the form of the posterior distribution. For instance, reference texts such as Carlin & Louis (2001), Gelman et al. (1995), and Gill (2002) present derivations for Bayesian posteriors for numerous prior-likelihood combinations. However, posterior distributions can be very complicated. Additionally, in multidimensional models, marginal posterior distributions must be calculated for each variable of interest. Marginalization requires integration, and not all probability distributions can be analytically integrated (Carlin & Louis; Gelman et al.; Gill; Johnson & Albert, 1992; Robert & Casella, 1999).

A venerable maxim in research is ”Quit when the math gets messy.” Confronted by these mathematical obstacles, researchers did not quit. They innovated, and Monte Carlo techniques were their answer to the ”messy math” problem. Monte Carlo techniques entail computer simulation of posterior distributions. In some cases, the values from the target distribution can be generated directly, for statistical packages can now generate random variables from many distributions. However, in many other cases, values from the target distribution cannot be generated directly. In these cases, values are first generated from a distribution from which samples can be easily drawn. A distribution so used is called a *proposal* or *jumping* distribution and the values generated from it the *candidate* values. For instance, WINBUGS (Speigelhalter et al., 2003), an ”off the shelf” computer program for conducting MCMC simulations, often generates candidate values from a normal distribution with mean and variance equal to the mean and variance of the previous iteration’s estimates. A probabilistic decision is made to *accept* or *reject* each candidate value as a member of the posterior distribution.

Monte Carlo techniques are the tools for making this decision. MCMC modeling is distinguished from other Monte Carlo methods in that its output is a Markov chain of parameter values. A Markov chain is a set of values with a *memoryless* property (Gill, 2002). After a burn-in set of values, each chain value will, in theory, be statistically independent of all values in the chain except the one value immediately preceding it. Additionally, after this burn-in period, an MCMC simulation of the joint-posterior distribution of interest will have converged to this distribution (Gill; Johnson & Albert, 1999).

The Metropolis-Hastings algorithm and a special instance of it, the Gibbs sampler, are the core techniques in MCMC. Gibbs sampling is a technique for calculating marginal distributions when analytic calculation is not feasible. It is especially useful in high-dimensional problems (Gill, 2002). In the Gibbs sampler, one simulates values from each parameter's distribution conditional on the most recent estimates of all other parameters in the model. When this process is repeated a sufficiently large number of times, these conditional distributions converge to their respective marginals (Casella & George, 1992). How large a number of times needed is a practical matter depending on the model, the data, and the chosen sampling technique. Some Gibbs samplers converge within 100 iterations; others have taken over 1.5 million iterations.

The Metropolis-Hastings algorithm is helpful when one can not sample directly from the conditional distributions for the model parameters. In this algorithm, one first generates a candidate value from the proposal distribution. Symbolically, this is usually referred to as $q(x|x_{t-1})$, where $t-1$ is the number of the iteration just completed. One then calculates the probability that the candidate value belongs to the posterior distribution. Symbolically, this distribution is often referred to as $f(x_c|y)$, where c refers to candidate and y to the observed data. The assignment of a probability is done via the acceptance ratio. When the proposal distribution is symmetric, this acceptance ratio simplifies to the ratio of posterior density of this candidate value to the posterior density of the chain's current value for the parameter(s) in question (Equation 14). When the ratio is greater than 1, the candidate value is accepted with probability 1.0. When less than 1, the candidate value is accepted, if the acceptance ratio is greater than a generated random uniform deviate (Carlin & Louis, 2001; Chib & Greenberg, 1995; Gelman et al., 1995; Gill, 2002; Hastings, 1970; Johnson & Albert, 1999). The result of this pattern is that higher-density regions of the posterior are sampled more often (Gill, 2002).

$$p(Accept) = min(1, \frac{f(x^c)}{f(x^{t-1})} \tag{14}$$

Caveat Lector

The expectation is that Markov chains will yield correct posterior distributions. However, there is fine print on the MCMC contract. If the researcher does not pay sufficient attention to the convergence and mixing of a chain, the distribution received may be something other than the distribution expected. Statisticians list several mathematical properties considered vital for understanding convergence. First, for a converged chain, the marginal distribution for the parameter of interest remains the same even as more iterations are included. This characteristic is referred to as *stationarity* (Gill, 2002).

Technically, a chain must meet three conditions in order to converge to its stationary distribution. It must be *irreducible*, *aperiodic*, and *positive recurrent* (Roberts, 1996). *Irreducibility* means that all values in the chain can, in theory, be transitioned to from any other value in the chain regardless of the starting values. A reducible chain is the mathematical equivalent of the blind alley in a maze. *Recurrence* means that the chain can, in theory, revisit any sampled value later in the chain. *Positive recurrence* means that the expected time to return to a given value is less than infinity (Gill, 2002; Robert & Casella, 1999; Roberts, 1996; Tierney, 1994, 1996). Those who would develop new MCMC algorithms must prove that the resulting stationary distribution is the posterior distribution of interest and must demonstrate the properties of irreducibility, aperiodicity, and recurrence or their necessary and sufficient conditions. Proofs concerning these attributes can be found in Robert & Casella and in Tierney (1994).

Mixing is the second concern. After convergence, the chain should move randomly, or nearly so, through points in the posterior distribution. If a chain is mixing well, one can be more certain that the representation of the posterior distribution is accurate. One of the signs that a chain is able to visit all of the points in the posterior distribution is that chain values quickly lose their correlation with each other over time. This property is referred to as low autocorrelation. Details can be found in Robert & Casella (1999).

Aids to assess convergence and mixing abound (Cowles & Carlin, 1996; Gelman et al., 1992; Gill, 2002). The sampling history plot is the first resort. This tool plots chain values for a given parameter on the y axis and time-point values on the x axis. A *beeline* from the initial chain value to a center of activity suggests convergence. Figure 17 shows the history plot for two chains started from different initial values, but making a bee-line for the same area of the posterior distribution (Speigelhalter et al., 2003). *Wandering* suggests nonconvergence (Figure 18), whereas *snaking* (Figure 19) suggests high autocorrelations (Gill). The ideal sampling history shows random fluctuations around a central point (Figure 20).

Gewecke's (1992) time sequence diagnostic and the Gelman & Rubin (1992) multiple sequence are frequently employed to assess convergence. Gewecke's test is essentially a Z-test comparing the marginal mean of a parameter earlier in the chain and its mean later in the chain. If the chain has reached its stationary distribution, the disparity between these means should be small (Gill, 2002). Technical details for the Gewecke test appear in Chapter 4.

The Gelman & Rubin statistic is essentially an ANOVA comparing between-chain variance and within-chain variance (Gill, 2002). If a Markov chain possesses all of the mathematical properties described above, then several chains starting from different initial points should have the same stationary distribution. One expects between-chain variance to be small compared to within-chain variance. Finally, a mathematical result of positivity and aperiodicity is that the between-parameter correlations are 0 in the limit (Robert & Casella, 1999). High unmodeled interparameter correlations suggest nonconvergence (Gill). Technical details for the Gelman & Rubin statistic appear in Chapter 4.

Mixing can be assessed by a parameter's autocorrelations. If values for a parameter remain correlated despite a long chain, it is likely the sampled values do not represent the full posterior distribution (Gill, 2002). For practical purposes, researchers permit themselves a lag of 40-50 iterations for the within-parameter correlations to go to 0. Technical details for calculating the autocorrelations appear in Chapter 4.

Practical Benefits and Drawbacks of MCMC Estimation

Whether one agrees or disagrees with the Bayesian philosophy, the Bayesian setup is the only way to obtain parameter estimates for high-dimensional and complicated functions such as the GMIRT model. In addition, MCMC estimation offers several other benefits. First, in Bayesian estimation, it is no longer a problem to obtain ability estimates for examinees who answer all questions right or all questions wrong . Because of the influence of the priors, finite posterior ability estimates can be achieved. Similarly, likelihoods that are not statistically identifiable can be combined with priors to produce unique posterior distributions (Johnson & Albert, 1999; Hambleton & Swaminathan, 1985). [3] As an additional benefit, MCMC estimation can also handle several likelihoods in one analysis. For instance, Patz & Junker (1999b) incorporate dichotomous and polytomous items in the same analysis.

However, Bayesian estimation in general and MCMC present numerous dangers along with the good fruits. First, Bayesian estimates are statistically biased (Lord, 1980). If the priors are quite different from the true posterior parameter values and the sample size is small in relation to complexity of the model, serious bias can result. Concerning a five-dimensional instantiation of the three-parameter multidimensional normal ogive model, Beguin & Glas (2001) reported that an extremely misspecified prior (true $\mu = 0$, prior μ specified as 1) with a sample size of 2000 examinees produced unfavorable estimates for the $\mathbf{a}$-parameters. Additionally, Bayesian estimates are not invariant to transformations, as are ML estimates (Lord, 1980).

[3] The use of the term *identifiable* for these models is highly controversial. Some authors readily use it (e.g., Johnson & Albert), whereas others shun it (e.g., Gelman et al., 1995).

Finally, MCMC estimation is exceptionally intensive with regard to computational resources (Bolt & Lall, 2003). On a Pentium IV 3.06 GHZ machine with 168MB of RAM, the GMIRT model in two-dimensions with 30 items and 4,000 examinees completed, on average, 1,000 iterations every 24 hours. Even with the very best in desktop computing of 2004, it is not uncommon that researchers must run their MCMC chains for several days or weeks for complex models. This investment of time is a serious impediment to the use of MCMC for IRT estimation in applied settings, especially in the for-profit sector.

Technical Aspects of MCMC and IRT

To recapitulate, unidimensional dichotomous and polytomous models, multidimensional compensatory normal ogive, compensatory logistic dichotomous and polytomous, normal compensatory , noncompensatory Rasch (MLTM), and variable compensation models with constraints (Ackerman & Turner, 2003; Beguin & Glas, 2001; Bolt & Lall, 2003; Cohen & Bolt, 2002; Jones & Nediak, 2000; Patz & Junker, 1999a,199b;Seagall, 2002) have been estimated successfully with MCMC methods. In these studies, authors have used either the ”off the shelf” software, WINBUGS, prefabricated routines from the C/C++ class library, MCMC-PAK, or created their own C/C++ or FORTRAN programs using the main algorithms described above.

Although created in a biostatistics environment, WINBUGS can operate on an exceptionally wide variety of models and enjoys serious use among psychometricians (Ackerman & Turner, 2003; Cohen & Bolt, 2002; Cohen, Wollack, Bolt, & Mroch, 2002). In some cases, it has been possible to speed WINBUGS' estimation by fixing a subset of parameters to estimates from an commercial MLE-based program (Cohen & Bolt). Technical details of WINBUGS' operation may be found in Gilks, Best, and Tan (1995), Gilks, Thomas, and Spiegelhalter (1994), Gilks and Wild (1992) and Spiegelhalter et al. (2003).

MCMC-PAK is a set of C/C++ routines created from a foundation of numerical

C/C++ classes (Martin & Quinn, 2003). The routines were created to bypass the R/S-PLUS languages' inability to rapidly execute loops. MCMC-PAK offers a routine for compensatory multidimensional IRT, but each subject's abilities are presumed uncorrelated. The main drawback of MCMC-PAK is that it does not offer the same flexibility in specifying models as WINBUGS.

MCMC-PAK's routine for compensatory multidimensional IRT employs data augmentation (Johnson & Albert, 1992; Tanner & Wong, 1987). For IRT models, this involves putting a hyperparameter on the item thresholds, namely creating for each person and item a normally distributed variable with mean = $\sum_{j=1}^{k} a_{ik}\theta_{pk} + d_i$ and standard deviation = 1. A very great benefit of using data augmentation for compensatory IRT models is that the full conditional distributions for the other variables are readily calculated and can have samples easily drawn from them. Hence, direct Gibbs sampling may be employed. Details on implementing data augmentation in IRT can be found in Johnson & Albert, 1992 and Beguin and Glas, 2001. It does not appear that there have been simulation studies assessing the performance of MCMC-PAK's routine for estimating multidimensional IRT models.

Many authors have written their own MCMC estimation routines with great success. Béguin and Glas (2001), working in FORTRAN, utilized direct Gibbs sampling and the data augmentation to estimate the parameters of the three-parameter multidimensional normal ogive model. Patz and Junker (1999a,1999b), working in S-PLUS, employed Metropolis-Hastings within Gibbs to estimate the parameters of unidimensional dichotomous and polytomous logistic models.

Writing custom MCMC programs gives the author ”ultimate flexibility, but also ultimate responsibility” (Gill, July 2003, personal communication). When writing their own MCMC programs, authors can group parameters as needed to reduce between-parameter correlations, select and modify jumping distributions, reparameterize models to foster better movement through the target distribution (see Robert & Casella, 1999) and, in general, do whatever it takes to estimate a particular model more efficiently. However, the authors must have a thorough knowledge of the mathematical basis of MCMC modeling and conditional probability in order to insure that their program does, in fact, run an MCMC chain and outputs the correct posterior distribution. To date, there appear to have been no studies comparing the efficacy of MCMC-PAK routines, WINBUGS, and ”hand-programmed” routines. Authors have, however, been very careful to report the results of simulation studies assessing their own programs' performance.

Other Methods of Assessing Noncompensatory Processing

Researchers were interested in noncompensatory processing well before IRT gained ascendancy in the measurement world. The non-IRT methods of assessing noncompensatory processing include regression, ANOVA, and laboratory assessment of behaviors thought to indicate noncompensatory processing. Regression-based assessment of conjunctive processes usually involves the conjunctive utility function (Equation 15).

$$U_{pi} = \Pi_{k=1}^{m} x_{p_{ik}}^{a_{ik}} \tag{15}$$

U_{pi} represents the utility of a stimulus for a person, each $X_{p_{ik}}$ a person's observed score on a component/dimension, k for stimulus i, and $a_i k$ the estimated weight of that component. Utility is a numeric representation of the value of a stimulus for the decision maker. A rater using a conjunctive method will assign high utilities to stimuli having no low scores on any dimension. For instance, a camper using a conjunctive decision rule to select items to take on the trip would select only those items light enough *and* useful enough. To estimate the component's a_i- parameter, each side of the equation is transformed by its natural logarithm, and then ordinary least squares regression is used to estimate the log of the weights. This regression technique is the most widely used method of assessing noncompensatory processing (See Einhorn, 1969, 1970, 1971; Ganzach & Czaczkes, 1995).

Conjunctive processing may also be viewed as an interaction between component abilities. ANOVA and linear regression with higher-order polynomial terms lend themselves to assessing noncompensatory processing in this framework (Wiggins & Hoffman, 1968; Slovic, 1969). *Process tracing* is the predominant laboratory method for assessing noncompensatory processing (Payne, 1976). In process tracing, participants search through information provided about a hypothetical decision they are asked to make. The participants' patterns of search through the information is analyzed. In particular, participants are observed to see whether they examine the same amount of information per stimulus or a variable amount, and whether they search the provided information within a stimulus or across stimuli. Searching the same amount of information for the stimuli is thought to be consistent with a compensatory strategy. Searching a variable amount of information per stimulus is thought to be consistent with a conjunctive strategy. The variable search pattern can be thought of as searching information about a stimulus until a disqualifying score is found. Searching between dimensions on a stimulus is thought to be consistent with a compensatory strategy, whereas searching within a dimension across stimuli is thought to be consistent with a conjunctive strategy. Again, one can think of a decision maker as trying to take a ”mental average” or ”weed out” stimuli. *Talk-aloud* protocols are also frequently used (Payne; Newel & Simon, 1972). In a *talk-aloud* protocol, the participant is asked to describe each step in his or her decision-making process to the experimenter.

Cognitive Theory for Inductive Reasoning Tasks

Induction is the development of rules from a set of specific instances (Pellegrino, 1985, p. 195). For instance, in a number series problem, an examinee views the first few members of a series and then must draw inferences about what rule or rules determine the series (e.g., 5,10,15,20 ... answer = 25). For the past 100 years, induction has been considered central to human intelligence (Embretson, 2002). Spearman (1927) described two important components of the general factor of intelligence, *g*, to be *eduction of relations* (rule induction) and the *eduction of correlates* (analogy), and Thurstone included induction as one of his primary mental abilities (Thurstone, 1938, cited in Pellegrino). Inductive reasoning tasks have high correlations despite their different subject matter content (Carpenter et al., 1990). Moreover, in reanalyses of data from well known factor-analytic studies of intelligence, inductive reasoning tasks showed either very strong loadings on *g* or very strong loadings on *gF* (fluid intelligence) (Carroll, 1993).

Psychometric tasks thought to tap inductive reasoning ability include series completion problems, verbal and geometric analogies, and matrix completion problems. One of the main selling points of most of these tasks for intelligence testing is that they are thought to minimize the role of learned experience (Carpenter et al., 1990). In series completion problems, the examinee must induce the pattern in a series of numbers or letters and then demonstrate understanding by determining the next number or letter in the series. In verbal or geometric analogies, the examinee is presented with two words or objects sharing a relationship. He or she must then induce the relationship between the first two objects, induce the relationship and then demonstrate understanding by determining the fourth element in the analogy (Thurstone, 1938; Pellegrino, 1985; Sternberg, 1977). For instance, in the verbal analogy, *horse* is to *zebra* as *pig* is to . . ., the correct answer among *giraffe*, *antelope*, *boar*, and *cow* is *boar*. Sternberg (1977) and Mulholland, Pellegrino & Glaser (1980) provide cognitive task analyses of analogies. Embretson and her colleagues (Embretson, Schneider & Roth, 1986; Whitely, 1981; Whitely & Schneider, 1980) have incorporated these task analyses into IRT modeling of analogy items.

In matrix completion problems, the examinee is shown an incomplete set of combinations of geometric figures. The examinee must induce the relationship between elements of the figures and demonstrate understanding by generating the missing figure or selecting it from a list of choices. The Raven's progressive matrices (Raven, 1962) and the matrix subtest of the Culture-Free Intelligence Test (Cattell, 1940) are the best known matrix completion problems. Cognitive psychologists and psychometricians have paid substantial attention over the past decades to the analysis of matrix completion tasks (Carpenter et al., 1990; Embretson, 1998, 2002; Hornke, 2002; Hornke & Habon, 1986; Hunt, 1974; Jacobs & Vanderwerter, 1972; Ward & Fitzpatrick, 1973).

Carpenter et al.'s (1990) taxonomy is the best known and most used of these cognitive task analyses. The taxonomy was developed on the Raven's Advanced Progressive Matrices(Raven, 1962). Additionally, this taxonomy has been adopted by other psychometricians in studies of online item generation (Embretson, 2002). The language of Evidence Centered Design (ECD; Mislevy et al., in press) can simplify discussion of the details of this taxonomy. In ECD, variables that may affect performance on a task are divided into *student model* variables and *task model* variables. *Student model* variables are attributes within the student, and *task model* variables are features of the tasks. For instance, vocabulary knowledge in a reading test would be a student model variable. On the other hand, sentence length would be a task model. Carpenter et al.'s taxonomy describes not only examinees' cognitive processes for solving the matrices but also the solution rules built into the matrices themselves.

Carpenter et al. (1990) described two student models, FAIRAVEN (Figure 21) for less-able examinees and BETTERAVEN (Figure 22) for more able examinees. From the diagrams, one can clearly see that both models are highly complex exhibiting the fine grain typical of cognitive task analyses. *Correspondence finding* refers to an examinee's ability to group objects for action by a rule. For instance, in BAS matrix 17, the examinee might realize that *shape* changes and *number* changes. *Row-wise rule induction* and *generalization* refers to the examinee's formation of a rule based on his or her observed correspondences. For instance, in item 17 , the examinee might articulate, ”lines increase in number by 1 in each frame.”

The main difference between the two models, is that executive control processes (i.e, the goal monitor) are the main component of the model for the more able students. Carpenter et al. (1990) hypothesized that better performing examinees are better able to maintain order in their problem solving activities. For practical purposes, psychometricians often attenuate this model to include only *correspondence finding* and *control processes*. Evidence suggests that these two features are closely related to item difficulty (Embretson, 2002).

The task model for the Ravens Progressive Matrices comprises the rules hypothesized to govern item solution. Table 3 lists Carpenter et al.’s (1990) rules along with the names by which some appear in other researchers’ taxonomies. *Constant in a row* refers to an element being repeated in each frame of a row. Other researchers have named this rule *identity relations* and permitted it to apply to rows or columns (Hornke, 2002; Hornke & Habon, 1986; Jacobs & Vandewenter, 1972; Ward & Fitzpatrick, 1973). Item 1 of the BAS matrices shows identity within a row; each shape name repeats within a row. *Quantitative pairwise progression* refers to an increment or decrement in some attribute of adjacent frames. For instance, in item 15 of the BAS matrices, the cluster of shapes rotates 90 degrees from frame to frame within each row. *Figure addition* refers to situations in which elements in frames 1 and 2 are superimposed make a particular element in frame 3. *Figure subtraction* refers to situations in which elements of frame 1 when subtracted from elements of frame 2 form a particular element of frame 3. BAS matrix 7 shows figure addition; the addition of a horizontal line makes each figure complete. BAS matrix item 16 shows figure subtraction; the removal of the circumscribed shape makes each figure complete.

In *distribution of 3 values*, an attribute takes on one of three possible values, and each of these values must appear in a different frame in each of 3 rows of the matrix. *Distribution of 3 values* has also been called the *latin square* rule (Ward & Fitzpatrick, 1973). Item 5 of the

BAS matrices is an example of *distribution of 3 values*. Over the life of the matrix, each shape name appears once in the first column, once in the second, and once in the third. In the present investigation, the term *rule of three* is used in place of the term *distribution of 3 values*. *Distribution of 2 values* is analogous to an incomplete block design; one of the frames in the matrix does not receive any value of the attribute. In this study, the term *rule of two* is used instead.

Empirical support is available for Carpenter et al.'s (1990) taxonomy. The authors developed computer simulation models incorporating some or all of the rules in their taxonomy. The computer simulation programs performed similarly to human examinees with regard to errors. For instance, BETTERRAVEN, programmed to recognizes more rules of the taxonomy than FAIRAVEN, made almost no errors. On the other hand, FAIRAVEN attained a total score identical with the median score of a group of human participants.

Working Memory: Background Information

The distinction between consciousness and memory reaches several hundred years into the past of western philosophy and psychology. The core concept remains the same, but the names have changed somewhat over the years. For instance, John Locke referred to the faculties of *contemplation* and *memory* (Richardson, 1996). In 1890, William James referred to the distinction between *primary* and *secondary* memory (Hunt & Ellis, 1993). In the 1968, Atkinson and Shiffrin named these components, *short-* and *long-term* memory. Today, we largely refer to the distinction between *working* and *long-term* memory.

There are several main theoretical viewpoints on working memory. In-depth treatment of the common ground among and differences between the theories can be found in Logie (1996), Richardson (1996) and, Shah and Miyake (1999). Some researchers view working memory as those contents of long-term memory that are activated (Cowen, 1999). Activation refers to a stored piece of information that is in consciousness or readily available to it. The multicomponent model separates working memory into domain-specific storage components (one for visual information, one for auditory information) and into a central executive, a component that directs processing and coordinates information from the other components (Baddely & Logie, 1999). As mentioned earlier, there is also a school of thought that emphasizes the role of controlled versus automatic cognitive processes (Ackerman, P., 1988; Hasher & Zacks, 1979; Engle et al., 1999). Controlled processes require working memory capacity, whereas automatic processes do not. Controlled processes are similar to the central executive in the multicomponent model.

The theory selected as the *lens* for this study of noncompensatory processes is the theory of automatic versus controlled mental processes. It was selected because of its holistic nature and parsimony. Additionally, it fits nicely with the idea that expertise is the result of automatization of processes.

More About Controlled Processes/Executive Processes

Because controlled and executive processes are central to this study, more discussion of them is given. It is generally agreed that controlled processes and executive processes control attention, maintain information and goals, and inhibit extraneous information (Miyake, Friedman, Emerson, Witzki, Howerter, & Weger, 2000; Engle et al., 1999). However, the degree of overlap between controlled and executive processes is an unsettled question. Some researchers (Engle et al.) would view the overlap as total. Others would not. Miyake et al. found that the some dual task activities combining spatial and verbal tasks failed to load on factors they associated with executive functioning. It could well be that there is some element of executive processing that is associated with success of tasks such as the Ravens and BAS matrices not fully explained by controlled processes.

Development of Working Memory in Childhood

It is well known that children’s performance on working memory tasks improves steadily with age (Schneider & Pressley, 1997). Most research suggests that it is not children’s working memory capacity that increases but rather their efficiency in using knowledge and strategies (Flavell, 1993). In classic studies, Chi (1978) found that child chess experts could retain a greater number of chess positions in working memory than adult chess novices, and Case, Kurland, and Goldberg (1982) found that an adult’s working memory capacity could be reduced to that of a six-year old by making nonsense words the stimuli to be recalled.

Working Memory and Test Design

Working memory has long been an important construct investigated in test design. Increased working memory loads in test items have been consistently related to higher item difficulty (Embretson,1998, 2002; Junker & Sitsma, 2001; Kyllonen, 2003). For instance, working memory scores for matrix completion items designed by Embretson (2002) correlated .74 with their Rasch item difficulty.

Working Memory Scores for BAS Matrices

In the current study, Carpenter et al.’s (1990) *task* model provided the basis for a working memory score for each BAS matrix item. Embretson (1998,2002) used Carpenter, Just & Shell’s (1990) taxonomy to assign working memory scores to matrix completion items. The rules in Carpenter et al.’s hierarchy received numeric ratings according to their place in the hierarchy, the *identity* rule receiving a 1, and the *distribution of 2 values* a 5. To obtain a final working memory score, the scores for each rule required by an item can be summed or weighted by rule difficulty and then summed. Details on the adaptation of Carpenter et al.’s taxonomy to the BAS matrices are given in Chapter 3.

Past Research Comparing Compensatory and Noncompensatory Processes

Little research has been done assessing the performance of noncompensatory models in tasks of interest in educational assessment. Among the studies to date, none has found evidence for superiority of a noncompensatory model to a compensatory one. Van Leeuwe & Roskam (1991) found that a noncompensatory multidimensional model provided worse fit to LSAT data than a compensatory model. In a recent study, Bolt and Lall (2003) found that a two-dimensional compensatory two-parameter logistic model provided a better fit to data from a test of English usage than a multidimensional Rasch model. However, Bolt and Lall conceded that the improved fit may be due to the addition of discrimination parameters in the two-parameter compensatory model. Although no one can really know what the implications of these early results are, Mislevy et al. (2001) found that a conjunctive model resulted fairly low reduction in posterior variance, a fact suggesting the model and data did not suit each other. In a simulation study involving the reliability of cut scores, Frye (2001) found that conjunctive scale scoring gave worse results than many other methods. Finally, when noncompensatory processes are represented by polynomial or interaction terms in a general linear model, very few researchers ever found indications that such terms were statistically significant or fit the data better (Goldberg, 1968,1971; Hammond & Summers, 1965; Hoffman, 1960; Slovic, 1969).

If people occasionally use their cognitive skills in a noncompensatory fashion or tasks demand such a configuration, then why does the compensatory model perform so well? The likely explanation rests in the moderate positive correlation between cognitive traits (Spray & Ackerman, 1986; Spearman, 1904, cited in Paik, 1998). In other words, it is far easier to discover an individual whose performance can be fit by both models than it is to discover one whose performance is fit by one and not the other. The situation is quite similar to the good fit shown by the simple linear model even in the presence of interactions (Yntema & Torgerson, 1961).

If the compensatory model performs so well, then why do people persist in chasing after the noncompensatory model? First, the compensatory model may not be a good fit for all individuals in a sample or for all tasks in an item-pool. In large-scale testing, these misfitting persons may amount to several thousand individuals. Perhaps their future achievement is being poorly predicted by compensatory models. Identifying these individuals and tasks could help shield testing organizations from lawsuits. Finally, the compensatory model may not be an accurate description of the task demands or how people are combining their skills.

Working Memory and Noncompensatory Processing- Past Research

Past research on the relationship between the working memory and noncompensatory processing mostly concerns decision making. Results have been mixed. Studies in which the number of choices available in a decision-making task (e.g., number of apartments in an apartment selection task) were varied have generally shown that subjects show more noncompensatory processing as the number of choices increases (Billings & Marcus, 1983; Olshavsky, 1979; Payne, 1976). Most researchers theorized that subjects resorted to conjunctive cutoff rules as a simplifying device. However, studies, in which the amount of information available for each choice (e.g., attributes for the apartments) was varied, have generally not shown differential noncompensatory processing as a function of information load (Einhorn, 1971; Payne, 1976). On the other hand, Oglivie & Schmitt (1979) presented anomalous results of increased compensatory processing when the amount of information to consider about each alternative was small. Researchers do not appear to have publicly speculated as to why different foci of working memory demands can lead to different results.

Although not investigating compensatory processing per se, Mulholland, Pellgegrino and Glaser (1979) found that a nonadditive model for the relationship between number of elements and number of spatial transformations in geometric analogies fit data for both errors and response times better than a simple additive model. Mulholland et al. attributed the increase in non-linear processing to mental resources' diversion to working memory as items became more complex.

CHAPTER III

METHODOLOGY

Overview of Chapter

This chapter presents the details on how both the simulation and application studies will be conducted.

Simulation Study Design

A Rasch model was selected for the present study for two reasons. First, it is often prudent to start with the simplest instance of a model before proceeding to more complicated instances. Secondly, as will become clear later, the Q-matrix for the substantive model weighs correspondence finding and executive control processes equally for all items. The selected levels of sample size, item pool size, and interdimensional correlation were fully crossed for a 2 x 2 x 3 design. One replication was performed per condition. The number of ability dimensions was held constant at two. Table 4 shows the levels selected for sample size, item pool size, and interdimensional correlations.

S-PLUS version 6 (Insightful Corporation, 2001) and FORTRAN routines were used to generate data according to the GMIRT Rasch model. One set of item parameters was generated, and new examinee data was generated for each condition. The $\mu_{\mathbf{c}}$ parameter was generated according a beta distribution with shape parameters 2 and 5. The mean of this β distribution was .2857, reasonably close to the .15 estimated for the general μ_c parameter estimated in Ackerman & Turner (2003). Discrimination parameters were fixed at 1.0. The $\mathbf{b}$-parameters were generated according to separate random normal distributions with a mean of 0 and variance of 1. The generated item parameter values appear in the leftmost columns of all tables describing item parameter recovery (e.g., tables 39 and 41).

Examinee ability scores were generated according to the multivariate normal distribution with a mean of 0 and a standard deviation of 1 and a correlation appropriate to the experimental condition. Sampling error was introduced into the responses by generating a uniform deviate between 0 and 1 for each item for each person. When the model- predicted probability is greater than this deviate, a correct response is generated. When the model-predicted probability is less than this deviate, an incorrect response is generated.

Application Study Design

Student and Task Models for the BAS Matrices

The BAS matrices fit very well with Carpenter et al.'s (1990) decomposition of the Raven's Progressive Matrices. There are many more processing components in Carpenter, Just & Shell's BETTERAVEN model than can be accommodated by a continuous latent ability model. Limiting the detail in a statistical model is a challenge. One strategy is to limit a model to those abilities that have been theorized and/or shown to drive individual differences in responses to the item. Executive control processes and correspondence finding are hypothesized to drive individual differences for the Raven's Progressive Matrices (Carpenter et al., 1990; Embretson, 2002). There is much converging support for this idea. First, Carpenter et al. found high correlations between performance on the Advanced Raven's Progressive Matrices and the Tower of Hanoi problem. Both tasks are thought to depend on executive control processes. Second, Embretson (2002) found that the degree of abstraction required by an item's correspondences and the number of rules required strongly predicted item difficulty in the Advanced Raven's Progressive Matrices, as well as in a set of abstract reasoning items generated from Carpenter et al.'s model. Given these details and similarity in task decomposition of the Raven's and the BAS matrices, it is plausible to limit the student model for the BAS matrices correspondence finding and executive control processes.

The final Q-matrix for the BAS matrices weighs both abilities equally, hence every entry is a 1. Fortunately, this makes the GMIRT Rasch model theoretically appropriate to use. Table 5 details the decomposition of the BAS matrices items into Carpenter, Just & Shell's (1990) rules, our task model. The decomposition of nearly all of the items is simple and straightforward. In order to give more detail where possible, the terms *subtraction* and *addition* are used instead of the more general term *figure addition or subtraction*.

A handful of BAS matrices items do not readily fit with Carpenter et al.'s (1990) taxonomy (items 3, 4, 23, 28). Two of these can be decomposed with rules drawn from Jacobs & Vandewerter's (1973) taxonomy. Matrix 4 could, in theory, be decomposed as an identity matrix along its columns. However, the *flip over* rule (Jacobs & Vandewerter) is a more sensible choice. In the *flip over* rule, mirror images of column or row entries appear in the next row or column. Use of a *flip over* rule would permit the examinee to read along rows instead of columns, a more natural act, even for children. Two other decompositions are plausible for Matrix 3. First, one could use a modified *rule of three*, namely a rule of two reading across rows, but this seems an unnecessarily complicated decomposition. Jacobs & Vanderwerter's *reversal* rule appears to be the simplest decomposition; the objects in one row simply reverse position for the next row. Items 23 and 28 fight classification into existing taxonomies. Item 23 appears to involve the sequential overlay of figures presented vertically in the first pane of the matrix. For now, this rule has been named *sequential overlay*. This type of mental transformation is quite common in geometric analogies. Item 28 involves a *distribution of 3 values* with one element repeated, an *on-on-off* pattern. For this study, this rule has been named *permutation with one repetition*.

Certainly, many of the BAS matrices admit of more than one possible decomposition. Whether or why examinees use alternative decompositions is a research question in its own right. However, we follow the basic hierarchy of rules that Carpenter et al. (1990) built into FAIRAVEN/BETTERAVEN computer programs. First, Ravens matrices should be solved across rows where possible. Second, Carpenter et al.'s model invokes the simpler rules first. In other words, a simulated examinee would try simpler decompositions before moving on to more complicated ones.

Embretson (1998, 2002) used Carpenter et al.'s (1990) taxonomy to assign weighted working memory scores to matrix completion items. She assigned rules in Carpenter et al.'s taxonomy numeric ratings according to their complexity, the *identity* rule receiving a 1, and the *distribution of 2 values* a 5. To obtain a final working memory score, the scores for each rule required by an item were sum. To use weighted working memory scores in the present study, *complexity* scores for the new rules needed to be created. The *flip over* rule and the *reversal rule* were assigned a complexity score of 2. The *sequential overlay* rule and *permutation with one repetition* were assigned a complexity score of 3. Junker & Sijtsma (2001), in modeling children's transitive reasoning, operationalized working memory load as the "number of premises" in an item, an unweighted working memory score. Both types of working memory scores were calculated and reported in the current study. Column 4 of Table 5 shows the number of rules required by each item, and column 5 shows the weighted-working memory scores. Some liberties were taken with assessing the number of rules for items 25 and 26. In each of these items, only one rule, the *rule of two* is involved, but since it must operate on four separate objects, the rule is counted as having four instances.

The BAS data allow further analyses on the validity of the working memory scores. Working memory load has been found to be an excellent predictor of item difficulty in several tasks, including matrix completion items (Embretson, 1998, 2002; Junker & Sitsma, 2001; Kyllonen, 2003). The correlation between number of rules required by an item and the item's *p* value was -.74 in the BCS70 first follow-up data for the BAS matrices. The correlation between weighted working memory scores and the *p* value was also -.74.

Additionally, participants in the BCS70 first follow-up also completed the digit span task of the BAS. Digit or letter span has been used for decades as an indicator of working memory capacity. In such tasks, the participant is asked to repeat an aurally presented set of numbers. The more numbers a participant can repeat, the higher his or her digit span score. Therefore, one should expect that the correlation between the recall of digits score and the total score on the BAS matrices to be sizable and positive. However, this correlation was only .25, a disappointing result. It could well be, however, that digit span and matrix problems tap different abilities within working memory. Such a conclusion is consistent with findings that single task span assessments often do not predict functioning on higher level cognitive tasks, whereas more the resource intensive, dual-task span assessments often do (Engle et al., 1999).

Data

Data for Matrices subtest of the BAS were taken from the ten-year follow-up from the 1970 British Cohort Study (BCS70). The British Cohort Study is a longitudinal study of the health and physical, social, and educational development of 17,198 children born in the United Kingdom in 1970. The National Birthday Trust Fund and the Royal College of Obstetricians and Gynaecologists sponsored the study at its inception. At the ten-year follow-up, 14,875 (*N*= 7713 male, *N*= 7162 female) children were involved. The mean age of the children was 10.13(.21) years. For the ten-year follow-up, participant identification was obtained through school rosters and recruitment via letters to parents. The most recent follow-up occurred in 1999. The British Centre for Longitudinal Studies and the UK Data Archive manage and house the data for the British Cohort Study. These organizations make these data freely available to academic institutions.

Measures

The BAS matrix completion task is one of the subscales of the British Ability Scales (Elliott, 1978). Based on Thurstone's (1938) theory of Primary Mental Abilities, the BAS was developed to offer differential diagnosis of cognitive abilities. All scales show acceptable internal consistency, but as of the time of the British Cohort Study's first follow-up in 1980, there was very little information on the inter-rater reliability, test-retest reliability and the overall validity of the measure (Embretson, 1985). The BAS was designed as an individual cognitive test, but in the BCS70 first follow up, the four subscales that were administered were administered in group sessions (University of Bristol, 1970). The matrix completion task consists of 28 items. The instructions to the child are to draw in the missing part of the ”pattern”. Note that the matrix completion test is not a multiple choice test as in the DAS (Elliott, 1990). Coefficient α for the BAS matrices in the BCS70 first follow-up data was .847 (*N* = 8704). The DETECT (Zhang & Stout, 1999) index suggests only one dimension for the data.

CHAPTER IV

RESULTS AND DISCUSSION: SIMULATION STUDIES

GMIRT Rasch Model- Free b_1, b_2

The priors for free parameters were set as follows. The ability parameters were assigned to bivariate normal hyperpriors. The means of these hyper priors were in turn assigned to normal priors with means = 0, and the covariance matrix of these hyperpriors were assigned to a Wishart distribution with a scale matrix of $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$. The $\mathbf{b}-$ parameters were assigned normal priors with mean 0 and variance 2. The $\mu_{\mathbf{c}}$ parameter was assigned a beta prior with shape parameters 2 and 5. To prevent the occasional "infinite result" ,which can occur with logistic likelihood functions (Spiegelhalter et al., 2003), the prior distributions for the $\mathbf{b}$ parameters were truncated at -3 and 3 and the prior distributions for $\boldsymbol{\theta}$-parameters at -5 and 5. Johnson and Albert (1999) and Patz and Junker (1999a, 1999b) provide examples of priors for multidimensional item response models.

In keeping with past practice for estimating multiple difficulty parameters per item, the $\mathbf{b}$- parameters for the first item were fixed at 0Bolt and Lall (2003). To make the entire first item an *anchor*, the μ_c-parameter for the first item was also fixed at 0.

All MCMC runs were conducted in WINBUGS version 1.4 (Spiegelhalter et al., 2002) with a group of 3GHZ Pentium 4 machines with a minimum of 1GB RAM and a maximum of 2GB RAM. Limitations on WINBUGS' capacity to store information became apparent when runs substantially more complex than those originally piloted were attempted. Because of these limitations, all conditions involving 50 items were omitted from the study, and only the first 100 θ pairs were monitored in all remaining conditions. Each chain was run for 30,000 iterations with 5,000 discarded as burn-in iterations. A sample of the WINBUGS code used in the current studies is included in the Appendix.

Convergence

Convergence and mixing were assessed by examination of sampling history plots, autocorrelation tables, Gewecke's (1992) statistic, and running mean plots. Figures 23, 24 and 25 show typical sampling history plots for item and ability estimates for the freely estimated GMIRT model. The plots suggest that convergence and mixing were adequate for most of the μ_c parameters, some of the b parameters, few of the θ parameters and none of the interability covariance parameters. The wandering in the τ_θ chains may have been responsible for the strange pulsation evident in the θ chains (e.g., Figure 25).

Gewecke's (1992) statistic, however, revealed potential convergence problems not evident in the sampling history plots. The Gewecke test is a sensitive comparison of parameter means from a designated first part of the chain with those from a designated final part(Equation 16):

$$Z = \frac{\bar{\Theta_1} - \bar{\Theta_2}}{\sqrt{\frac{s_1(0)}{n_1} + \frac{s_2(0)}{n_2}}}, \tag{16}$$

where n refers to the number of iterations in the selected portion, the majuscule Θ refers to any generic parameter, and s to the spectral density, a variance calculation often used in time series data because it can account for the correlation between time points (Congdon, 2003; Gill, 2002). All Gewecke tests in the current study used the recommended initial .10 iterations for the first window and the last .50 for the second window (Gewecke, 1992).

Tables 6 - 8 list the Gewecke statistics for each item parameter. Table 9 reports the proportion of items and inverse covariance terms in each condition failing the Gewecke test ($|Z| >= 1.96$). On average, approximately .67 of items failed the Gewecke criterion for the b_1

and b_2 parameters (minimum mean proportion = .375, maximum = 1.0). The picture was slightly better for the μ_c parameter where, on average, approximately .33 of items failed the Gewecke criterion (minimum = .083, maximum = .875). For both b and μ_c parameters, convergence usually improved with decreasing sample size (N=3,000) and improved with increasing interability correlation, $r(\theta_1, \theta_2) = .3$ and $r(\theta_1, \theta_2) = .6$. Nearly all items in all conditions failed the Gewecke criterion for the τ_θ parameter. No further analyses were done concerning this parameter.

Table 10 shows summary statistics for the Gewecke statistics for the 100 pairs of θ estimates monitored. Approximately .75 of the monitored persons failed the Gewecke criterion (minimum mean proportion= .50, maximum = 1.0).

Tables 11 - 13 show the lag-50 autocorrelations for each item parameter in each condition. Table 14 shows summary statistics for the lag-50 autocorrelations for the inverse covariance terms and item parameters by condition. Autocorrelations, it should be remembered, should be 0 if a chain has reached its stationary distribution (Robert & Casella, 1992). Again, a generous lag, usually 50 iterations, is allowed for this journey to 0. High autocorrelations signal poor convergence, poor mixing or both (Gill, 2002). Equation 17 gives the standard calculation for lag k autocorrelation (National Institute of Standards, 2003).

$$r_k = \frac{\sum_{i=1}^{N-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{N}(y_i - y)^2} \tag{17}$$

Data from the autocorrelations largely complement results from the sampling history plots. The lag-50 autocorrelations were near 0 for all of the μ_c parameters except for those in the N = 3,000, $r(\theta_1, \theta_2) = 0$ condition, suggesting adequate convergence and mixing for most of the μ_c parameters. Autocorrelations remained substantial for the b parameters in all conditions except for b_1 in the $N = 3,000$, $r = (\theta_1, \theta_2) = 0$ condition. The grand mean of the autocorrelations was .28 (minimum mean autocorrelation=.10, maximum mean autocorrelation=.54). As seen in the Gewecke results and sampling history plots, autocorrelations were lower in the smaller sample size conditions, $N = 3,000$, for both μ_c and b parameters.

Table 15 shows summary statistics for the lag-50 autocorrelations for $\hat{\theta}$. Surprisingly, given the Gewecke statistics for the θ estimates and their bizarre sampling histories, the autocorrelations for most conditions were not especially large (.10 - .17).

Running mean plots can explain some of the b and μ_c estimates' large Gewecke statistics. Figure 26 shows typical running mean plots for b parameters under small sample size. Figure 27 shows running mean plots for μ_c parameters typical for all conditions. The means for μ_c and the b parameters became quite stable after 15,000 iterations, a result suggesting that convergence problems in the smaller sample size may be alleviated by allowing a longer burn-in period in future studies. Furthermore, most of the differences apparent on the running mean plots are really fairly small.

The running mean plots for the b parameters for the larger sample size (N=6,000) show a much more serious problem, unlikely to be solved by more iterations(Figures 28 and 29). The meandering of the trace lines suggests ”dimension swapping”; In the two-dimensional version of this phenomenon, values appropriate to the posterior distribution of one difficulty parameter for an item find their way into the other difficulty parameter's chain. The same phenomenon can happen for ability parameters (Bolt & Lall, 2003). The reader can compare the running mean plots for $b_{2[25]}$ (Figure 28) and $b_{1[25]}$ (Figure 29) for a particularly striking example. In some cases when working in two dimensions, the researcher can prevent dimension swapping for a pair of parameters by placing ordinal constraints on the priors for a particular item so that the maximum value for one dimension's estimate is always less than the minimum value for the other dimension's estimate (Bolt & Lall, 2003).

In light of all of the evidence on convergence in these data, one must conclude that there were serious convergence difficulties for the inverse covariance terms and ability parameters for all conditions, as well as for the difficulty parameters for larger sample sizes. There were some convergence difficulties for the difficulty parameters for smaller sample sizes and for the compensation parameters.

Parameter Recovery

Table 16 shows the true parameters, WINBUGS estimates, signed biases and Root-Mean-Square Errors (RMSEs) for the b_1 parameter for each item for all conditions when the sample size was 3000. Table 17 shows this information when the sample size was 6000. Tables 18 and 19 show this information for the b_2 parameter and Tables 20 and 21 for the μ_c parameter. Table 22 shows the average signed bias and RMSE for all conditions for the b_1, b_2, and μ_c parameters. The RMSEs were fairly large for the b parameters (.56 to .76) — larger than most that Bolt and Lall (2003) found for the b parameters in the noncompensatory Rasch (Table 23). In the current study, bias was usually negative for the b parameters, but it is unclear why. RMSEs were acceptable for the μ_c, approximately .10 for a parameter that could take on values from 0 to 1.0. There were no clear effects for sample size or interability correlation on estimation of the b or μ_c parameters.

Table 24 shows the correlation between the true and estimated values of b_1, b_2, and μ_c for all conditions. Additionally Table 24 shows the cross-correlations, $r(b_1, \hat{b_2})$ and $r(b_2, \hat{b_1})$.

Very high ($>$ .90) correlations between a parameter and its estimate suggests good estimation.[4] The correlations between μ_c and $\hat{\mu_c}$ were between .74 and .85 for all conditions– acceptable but not excellent recovery, a result consistent with the RMSEs. Correlations between b_1 and $\hat{b_1}$ ranged from .47 to .69, indicating largely unacceptable recovery. The correlations between b_2 and $\hat{b_2}$ were surprisingly high (.78-.88) given the disappointing RMSEs.

[4] There does not seem to be a standard reference for this threshold.

The most surprising result in Table 24 was that the correlation between the b_2 parameters and $\hat{b_1}$ exceeded that between the b_1 parameters and the b_1 estimates. Lim (1993) observed similar behavior with the b parameters of the noncompensatory model. This mismatch of correlations was probably due to the dimension swapping evident in the convergence diagnostics.

Tables 25 and 26 show the mean and standard deviation of the proportion reduction from prior to posterior standard deviation for all conditions for the b_1, b_2, and μ_c parameters. The posterior standard deviation will always be equal to or less than the prior standard deviation. Smaller posterior standard deviations reflect more accurate estimation of the the parameter in question (Carlin & Louis, 2000). Posterior standard deviations for the item parameter estimates in the current study reflected a reduction of approximately 40 to 56 percent of the prior standard deviation.

Table 27 shows the average signed bias and RMSE for all conditions for the ability parameters and for the expected number correct (ENC) score (Equation 18):

$$ENC_p = \sum_{i=1}^{ni} T_i(\theta), \tag{18}$$

where $T(\theta)$ refers to the two-dimensional GMIRT model from Equation 9 and ni to the number of items. The RMSEs for the θ parameters and ENCs were rather large. The RMSEs for $\hat{\theta}$ ranged from .52 to .86, and those for the ENCs were were approximately 4 points on a 25 point scale. The bias in the $\hat{\theta}$ estimates was usually negative and that in the ENC positive. As with the item parameter estimates, there were no clear effects of sample size or interability correlation.

Table 28 shows the correlations between the true and estimated values of θ_1, θ_2, and the ENCs by condition. The correlations between θ and $\hat{\theta}$ values were greater than or equal to .67 in all but two conditions. By themselves, these correlations would represent acceptable parameter recovery. However, considering the RMSEs and the sizable cross-correlations (e.g., $r(\theta_1, \hat{\theta_2})$), parameter recovery for the ability parameters appears inadequate. Surprisingly, the correlations between the estimated and true ENCs were all greater than .90. This result suggests that, even though bias was substantial, the ENC estimates might still be usable for applications requiring only rank ordering of examinees.

Tables 29 and 30 show the mean and standard deviation of the proportion reduction from prior to posterior standard deviation for all conditions for the θ estimates. Reductions ranged from 39 to 46 percent of the prior standard deviation.

GMIRT Rasch Model- Constraint, $b_1 = b_2$

Convergence

Figures 30, 31 and 32 show typical sampling history plots for item and ability estimates for the freely estimated GMIRT model. As in the free estimation study, convergence appeared adequate for most of the μ_c parameters, some of the b parameters, few of the θ parameters and, none of the interability covariance parameters.

Gewecke (1992) Z statistics were again conducted. Tables 31 - 32 list the Gewecke statistics for each item parameter. Table 33 reports the proportion of items and inverse covariance terms in each condition failing the Gewecke test. Surprisingly, more items failed the Gewecke test for the $\bar{b}$ parameter than for the b parameters in the free estimation study. On average, approximately .78 of items failed the Gewecke criterion for the $\bar{b}$ parameters (minimum mean proportion = .12, maximum = 1.0). As in the free estimation study, Gewecke results for the μ_c parameter were better than those for the difficulty parameters. On average, approximately .42 of items failed the Gewecke criterion for the μ_c parameter (minimum = .04, maximum = .79). As in the free estimation study, convergence usually improved with decreasing sample size (N=3,000) and improved with increasing interability correlation for both b and μ_c parameters. Again, nearly all items in all conditions failed the Gewecke criterion concerning the τ_θ parameter.

Table 34 shows summary Gewecke results for $\hat{\theta}$. As in the free estimation study, approximately .75 of the persons failed the Gewecke criterion (minimum mean proportion = .15, maximum = .98).

Tables 35 - 36 show the lag-50 autocorrelations for each item parameter in each condition. Table 37 shows summary statistics for the lag-50 autocorrelations for the inverse covariance parameters and item parameters by condition. Autocorrelations for the difficulty and compensation parameters were much improved in the constrained estimation study. For the $\bar{b}$ parameters, the sampling histories hint that this improvement may be due to better mixing.

Table 38 shows summary statistics for the lag-50 autocorrelations for $\hat{\theta}$. Autocorrelations for the θ parameters in the constrained estimation study were similar to those in the free estimation study.

Figures 33 and 34 show running means plots for selected $\bar{b}$ and μ_c parameters. Nearly all of the trace lines for the $\bar{b}$ parameters became more stable by 10,000 iterations. It is unclear why there was a peak at approximately 20,000 iterations for some of the $\bar{b}$ trace lines. However, these peaks represent small differences in $\hat{\bar{b}}$ values. For the μ parameters, nearly all of the trace lines turned stable by 15,000 iterations. For constrained estimation, all of the convergence difficulties can likely be overcome by a longer burn-in period in future studies.

Parameter Recovery

With b_2 constrained to equal b_1, the target of recovery becomes the simple average of the true values of these parameters. Table 39 shows the true parameters, WINBUGS estimates, calculated biases and RMSEs for the b_1 parameter for all conditions when

$N = 3,000$. Table 40 shows the same information for $N = 6,000$. Tables 41 and 42 show this information for the μ_c parameter.

Table 43 shows the average signed bias and RMSE for all conditions for the $\bar{b}$ and μ_c parameters. Estimation was dramatically better with a single b parameter than with multiple b parameters. RMSEs for the $\bar{b}$ parameter ranged from .16 to .25. Although better than available estimates from noncompensatory models, they remain worse than estimates from compensatory models. For instance in a recent study, RMSEs for the compensatory model's difficulty parameter were always below .10 (Bolt & Lall, 2003). Results for the μ_c parameter were similar to those from the free estimation study.

Table 44 shows the correlation between the true and estimated values of $\bar{b}$ and μ_c for all conditions. The correlations for $\bar{b}$ were nearly perfect, a further piece of evidence suggesting very good parameter recovery. The correlations for μ_c were similar to those from the free estimation study.

Tables 45 and 46 show the mean and standard deviation of the percent reduction from prior to posterior standard deviation for the $\bar{b}$ and μ_c parameters. Posterior standard deviations for the $\bar{b}$ parameter represented nearly a 95 percent reduction of the prior standard deviation. Furthermore, the reductions in posterior standard deviations for the μ_c parameter appeared greater than those seen in the free estimation study. Also, it appeared that the posterior standard deviation for the $\hat{\mu_c}$ decreased slightly with increased interability correlation and sample size.

Tables 47 - 50 show signed bias, correlation, and standard deviation reduction information for the θ estimates. In all respects, the ability estimates were similar to those from the free estimation study.

Issues in the simulation studies

One of the more startling results, or absence of results, one might say, is the failure to find improved estimation with increased sample size and worsened estimation with increased interability correlation(Ackerman, Kelkar, Neustel, & Simpson, 2001a; Béguin & Glas, 2001; Bolt & Lall, 2003). Only the μ_c parameters' posterior standard deviations behaved in line with this prediction. In other conditions, there were no clear effects of sample size or interability correlation.

Perhaps estimation in the lower sample sizes and lower correlations was worse than it needed to be, or perhaps the less good convergence in these conditions degraded the quality of estimation. Further studies with longer burn-in periods can address this issue.

CHAPTER V

RESULTS AND DISCUSSION: APPLICATION STUDY

Setup

Only complete response patterns for the BAS matrices were included in the MCMC analyses (N = 8,704). It appears that these participants were more able than the omitted participants both in terms of total score and proportion of completed items answered correctly. The mean total score for examinees completing the subtest (N = 8,704) was 16.69(5.00), and the mean proportion of completed items answered correctly was .60(.18). The mean total score for examinees (N=2,788) completing at least 1 item but not all items was 11.15(4.32) and the mean proportion of completed items answered correctly was .42(.16). Of the 14,875 participants in the second followup, 3,388 did not contribute any data for the matrices subtest.

The 8,704 responses were divided into main and cross-validation datasets of 4,532 records each. Odd numbered records were sent into the main dataset (A), and even-numbered records were sent into the cross-validation dataset (B).

Given the convergence problems in the simulation studies, all chains in the application study were run for 50,000 iterations with 25,000 discarded as burn-in iterations. For both of the datasets, two simultaneous chains were run for each of the models. Again, the models were the GMIRT Rasch model, the noncompensatory Rasch model, and the multidimensional compensatory Rasch model. A single difficulty parameter was estimated in all models. Priors were the same as those in the simulation studies.

Convergence

Convergence was assessed individually each of the two chains associated with the estimation of the three models and two datasets (12 chains). Table 51 shows the proportion of items failing the Gewecke criterion for each parameter group for all chains. Table 52 shows the mean autocorrelations for each parameter for all chains.

Both chains for the GMIRT model in the main dataset (A) ran well. Figures 35, 36, and 37 show typical sampling histories for the two chains in this dataset. The sampling histories for the τ_θ parameters showed improvement from the simulation study's results. Between 25 and 50 percent of items failed the Gewecke statistic for the μ_c parameter for the main dataset, but running mean plots suggest the differences driving these statistics were very small (Figures 38 and 39). Autocorrelations were near 0 for all parameters in the first chains except for the τ_θ parameter. Although improved, convergence and mixing for the τ_θ parameters were still unsatisfactory.

However, the chains for the cross-validation dataset (B) in the GMIRT model exhibited problems with convergence and mixing. Although showing excellent convergence, the first GMIRT chain in the cross-validation dataset showed mediocre mixing with autocorrelations for the $\bar{b}$ parameters larger than one would like (Table 52). Two particular matrices appear to have caused any mixing problems for the μ_c parameters. The mean autocorrelation for the μ_c parameters appeared to be inflated by large autocorrelations for matrices 25 ($r_{50} = .26$) and 26 ($r_{50} = .41$). Only a small proportion of persons failed the Gewecke test for the θ parameters for these chains (.05 - .27).

The second GMIRT chain for the cross-validation dataset belonged in the ”accidents happen” category. A sudden dip in the sampling history of the inverse covariance terms seems to have spread into the sampling histories of many other parameters. Figure 40 shows the sampling history of the inverse-covariance matrix for this chain, and Figure 41 shows the unfortunate sampling histories of four μ_c parameters (The two histories near the bottom of the figure are nonproblematic). The ”renegade” chain was excluded from further analyses.

All chains for the noncompensatory model ran fairly well. Almost no $\bar{b}$ parameters failed the Gewecke test. However, autocorrelations were higher than desirable for the $\bar{b}$ parameters. Figure 42 shows typical sampling histories for the $\bar{b}$ parameters for the noncompensatory models. These histories do not show evidence of poor mixing or poor convergence. Perhaps the autocorrelations reveal slight problems with mixing that do not show up in the sampling histories. Small proportions of persons failed the Gewecke test for the θ parameters for the first dataset (.05 - .27). A sizable proportion of persons in the cross-validation datasets for the noncompensatory model failed the Gewecke test (.35 - .96), but running mean plots suggest the differences driving these tests were quite small (Figure 43). As in the simulation studies, convergence and mixing for the τ_θ parameters were poor as evidenced by the Gewecke results and the autocorrelations.

No d parameters in the compensatory model failed the Gewecke test, and their autocorrelations were very near 0. However, nearly all of the θ parameters failed the Gewecke criterion and showed double-digit autocorrelations. The running mean plots suggest that the differences driving the Gewecke statistics were substantial for many θ parameters (Figure 44). The θ parameters from the compensatory model were excluded from further analyses. As with the noncompensatory model, all of the τ_θ parameters failed. The τ_θ parameters were excluded from further analyses for all models.

The Brooks-Gelman-Rubin diagnostic (BGR; Brooks & Gelman, 1998), a modified MANOVA, assesses the convergence of two or more MCMC chains by comparing the between-chain variance/covariance with the within-chain variance/covariance. If a set of chains has converged to the same stationary distribution, the between-chain variance/covariance should be small. Roy’s greatest root is used to calculate the distance between the within-chain variance/covariance matrix and a weighted average of the within-

and between-chain variance/covariance matrices. This root approaches 1 when the between-variance-covariance matrix is null. For use in the BGR statistic, Roy's greatest root is referred to as the multivariate potential scale reduction factor (MPSRF; Brooks & Gelman). Therefore, MPSRF values near 1.0 suggest that the set of chains has converged to the same stationary distribution.

The BGR (Brooks & Gelman, 1998) diagnostic was used in the current study primarily to show model identifiability. For the GMIRT model, no MPSRF was calculated for the main dataset because of the second chain's severe convergence problems. The inverse-covariance parameters were omitted from all of the MPSRF calculations, and the ability parameters were omitted from MPSRF calculations involving the compensatory model.

Within those constraints, model identifiability appears to have been adequate for all but one condition, the cross-validation dataset for the noncompensatory model. Perhaps the convergence problems in this dataset were more serious than previously supposed for the noncompensatory model. The θ parameters were removed from this dataset and the BGR re-run. The MPSRF decreased from 10.30 to near 1.0. These θ parameters were excluded from further analysis. Table 53 shows the final MPSRFs for each model and dataset for which the diagnostic could be calculated.

A Final Note on Inter-Parameter Correlations

The μ_c and b parameters in the simulation study were generated as statistically independent. They were modeled as statistically independent in both the simulation and application studies. However, the μ_{c_i} and $\bar{b}_i$ parameters in the application study were correlated approximately .70 across iterations. This correlation did not seem to hurt chain mixing.

Parameter Estimates

For estimation purposes, individual chains passing convergence and mixing assessments were combined within a dataset and model. For instance in the main dataset for the GMIRT model, the 25,000 retained iterations of chain 1 and 25,000 retained iterations of chain 2 were combined to form a single 50,000-iteration dataset.

Item Parameters

Table 54 shows the posterior means for the difficulty parameters for all models, and Table 55 shows the posterior means for the compensation parameters. The difficulty parameters behaves as expected for all models and datasets with the earlier items being the least difficult and the later items the most difficult. Again, the compensatory model's difficulty parameters have opposite signs from those of the GMIRT and noncompensatory models. The correlation between item number and difficulty was approximately .94 for the GMIRT and noncompensatory models and -.94 for the compensatory model.

Table 56 shows a frequency distribution for the μ_c estimates for both datasets. Given their potential range of 0 to 1.0, the μ_c parameters seemed low. In both datasets, nearly two thirds of the items had compensation parameters less than .20. However, research suggests that μ_c parameters as small as .20 can consistently introduce differential item functioning (Ackerman & Bolt, 1995).

For both the difficulty and compensation parameters, informal cross-validation appeared excellent. For difficulty parameters, the correlations between the main and cross-validation dataset's estimates for the three models were .99 for the GMIRT model, .99 for the noncompensatory model and .99 for the compensatory model. For the compensation parameter, the correlation between the main and cross-validation dataset's estimates was .92.

The correlations between the difficulty estimates for the GMIRT model, noncompensatory model, and compensatory model were exceptionally high, approximately

-.99 for correlations involving the compensatory difficulty parameter and approximately .98 for the other models' difficulty parameters. These correlations suggest that any of the three models can adequately capture item difficulty for the BAS matrices.

Table 57 shows the posterior standard deviations for the difficulty parameters for each model and dataset. Table 58 shows the proportion reduction from prior to posterior standard deviations for the difficulty parameters for each model and dataset. The mean percent reduction ranged from a low of .93(.08) for the GMIRT datasets to a high of approximately .96(.02) for the noncompensatory and compensatory models. The reductions in standard deviations for the difficulty parameters were excellent for all models.

Easier items showed less proportion reduction in their standard deviations for all models and all datasets. This, again, suggests greater uncertainty regarding the difficulty parameters for the easier items. The correlation between a model's difficulty parameter and its proportion reduction in standard deviation ranged from -.60 to -.72 for the compensatory model and from .62 to .77 for the noncompensatory and GMIRT models. For the compensation parameter, this correlation ranged from .79 to .83. Perhaps, limiting the analyses to complete response vectors only led to this improved reduction in posterior standard deviations for the more difficult items' $\bar{b}$ parameters.

Table 59 shows the posterior standard deviations the μ_c parameter for both datasets, and table 60 shows the proportion reduction in standard deviations for the μ_c parameter. The mean proportion reduction was approximately .72(.20). This was less good than the reductions observed for the difficulty parameters, but still a reasonable amount. The reductions were less for higher μ_c parameters

It was originally hoped that item-fit statistics, in particular the z-fit statistic (Drasgow et al., 1995), could be computed. This computation was deemed to be an exercise in futility in the current study because of the relatively low number of examinees whose ability parameters could be monitored in WINBUGS version 1.4 (Spiegelhalter et al., 2003) and because of the poor quality of parameter recovery for the ability estimates in the simulation study. Furthermore, any fit statistic relying on estimated abilities would likely give suspect results. No item-fit statistics were conducted in the current study.

For now, we will have to have to rely on what simple reasoning can tell us about item-fit for the GMIRT, noncompensatory, and compensatory models, namely that the difficulty parameters behaved as predicted by theory and were very highly correlated across models. As will be shown below, the compensation parameters also, to some extent, behaved as predicted.

Ability Parameters

Convergence was acceptable for the first 100 ability parameters for the main and cross-validation datasets for the GMIRT model and the main dataset for the noncompensatory model. Table 61 shows the correlations between the means of the posterior distributions for the ability parameters for the GMIRT and noncompensatory models for the main dataset. The correlations within a parameter and across models were all very high ($>= .88$), suggesting that the GMIRT and the noncompensatory models performed similarly in estimating person attributes. However, the cross-correlations for ability parameter estimates were also very high. This is most probably due to an insidious form of dimension swapping in which parameter estimates for two dimensions fall back to the average between the dimensions (S. Jackman, personal communication, February 12, 2004).

Table 62 shows the mean posterior standard deviations for the ability estimates for the GMIRT and noncompensatory models. Table 63 shows the mean proportion reduction from prior to posterior standard deviations. Overall, reduction was poor and was worse for the GMIRT than for the noncompensatory model. Reduction for the cross-validation dataset was negative — a theoretical impossibility. In short, the reduction results suggest great uncertainty about the estimates already shown to have high RMSEs in the simulation study.

No person-fit statistics were conducted because of the suspect quality of the ability estimates.

Comparative Model Fit

It had been hoped to assess comparative model fit through Bayes factors. The Bayes factor is similar to a likelihood ratio, except it can accommodate prior information and need not compare nested models (Gill, 2002; Kass,1993; Kass & Raftery, 1995). It is the posterior odds ratio for model 1, $M_1|\mathbf{y}$, to model 2, $M_2|\mathbf{y}$, where M represents the likelihood function for a given model.

Equation 19 shows this ratio of posterior odds.

$$\frac{\frac{\pi(M_1|\mathbf{y})}{p(M_1)}}{\frac{\pi(M_2|\mathbf{y})}{p(M_2)}} \tag{19}$$

As shown in Equation 20, when the two models are assigned equivalent prior probabilities, the Bayes factor reduces to the common likelihood ratio (Gill, 2002). Values of the Bayes factor greater than 1 suggest support for model 1, whereas values of the Bayes factor near 0 suggest support for model 2 (Gill, 2002).

$$\frac{\pi(\mathbf{y}|M_1)}{\pi(\mathbf{y}|M_2)} \tag{20}$$

WINBUGS version 1.4 can usually provide the deviance for a model, $-2*$Likelihood. The likelihood for a model can be readily calculated from it. However, the deviance is unavailable when any of the prior distributions is censored, as was done for the difficulty and ability parameters in the current study (D. Spiegelhalter, personal communication, June 18, 2004). Bayes factors were not calculated in the current study.

Overall model fit was assessed in a much less complex way via comparisons of the

ENCs among the models for which ability estimates were available. Table 61 shows the correlations between the ENCs and the total scores for the GMIRT and noncompensatory models for the main dataset. These correlations were quite high, suggesting that both models could estimate the total score very well. Past research suggests that, at the test level, compensatory and noncompensatory models give very similar results (Ackerman et al., n.d.). The present research supports this conclusion. The maximum absolute value of the difference between ENCs for the GMIRT and noncompensatory models was less than two points for both the main and cross-validation datasets.

Correlations between the RMSEs for the ENCs and the absolute value of the difference between θ_1 and θ_2 were .56($p < .001$) and .45($p < .001$) for the GMIRT model for both datasets and .80($p < .001$) for the noncompensatory model for the main dataset. The effect was more pronounced for the noncompensatory data. It is tempting to conclude that model fit worsens as actual θ values diverge. However, θ parameters were not particularly well estimated in the simulation study, so one can not validly make this conclusion.

Contour plots of probabilities of correct response were created for all BAS matrices for the GMIRT, noncompensatory, and compensatory models, as well as for the differences in probabilities of correct response between the GMIRT and compensatory model. Item parameter estimates from dataset A and an idealized quadrature grid were used. The mean of the idealized ability distribution was 0 and its standard deviation 1. Therefore, all difficulty estimates were rescaled to z scores for this set of analyses. Figures 45,46, and 47 show examples of these plots for very easy items, moderately difficult items, and very difficult items. These plots should help readers visualize which individuals would most likely be affected by compensation effects for each item.

For all items and nearly all ability levels, the GMIRT model produced probabilities of correct response less than those of the compensatory model. The grand mean of differences between the probabilities of correct response to the BAS matrices under the GMIRT versus the compensatory model was -.07 (minimum average difference = -.02; maximum average difference = -.12). Only item 11, the item with the highest compensation parameter among the BAS matrices, gave an average difference less than -.10. In short, when speaking on average results, the GMIRT and compensatory models were quite similar even at the item level.

Noncompensatory and compensatory models give overall similar results because their differences in probabilities of correct reponse are usually concentrated where the examinees are not (Ackerman, n.d.). For the current data, we expect to find the greatest differences in probabilities of correct response for examinees who were offline outliers. An examination of the contour lines reveals this to be true for most items. An additional point of interest in the contour plots is that the location of the maximum difference contour shifts to the upper right of the Cartesian plane with increasing item difficulty. In fact, when one reaches the most difficult item (item 28), the contour of maximum difference involves online outliers. This effect was not evident in previously issued contour plots for the GMIRT model, because they showed items of average difficulty (Ackerman, n.d.; Ackerman & Bolt, 1995).

When speaking of extreme differences between component abilities, the GMIRT and compensatory models could be quite different at the item level. Over three fourths of the items had maximum differences in probability of correct response less than -.22. Finally as one would expect, maximum differences in probabilities of correct response were strongly correlated with the compensation parameter ($r = .88, p < .001$), i.e., items that were very nearly compensatory did not show much difference in probabilities of correct response compared to items calibrated under the purely compensatory model.

Working Memory Load and Compensation

Table 64 shows the posterior means for the compensation parameters for the main and cross-validation datasets alongside the unweighted and weighted working memory loads for each item. As expected, working memory load and item difficulty were highly correlated

(Kyllonen et al., 2003). The correlations between unweighted working memory load and the posterior means of the difficulty parameters were .73 ($p < .001$) for the main dataset and .72($p < .001$) for the cross-validation dataset. These correlations were .64($p < .001$) for the main dataset and .63($p < .001$) for the weighted working memory scores.

The correlations between the μ_c parameters and the two working memory scores for the BAS matrices were lower than expected. The correlation between the unweighted working memory scores and the μ_c posterior means was .24 ($p = .22$) for the main dataset and .18 ($p = .37$) for the cross-validation dataset. The correlation between the weighted working memory scores and the μ_c posterior means was -.05 ($p = .79$) for the main dataset and -.10 ($p = .62$) for the cross-validation dataset.

Figures 48 - 51 show scatterplots for these four relationships. It quickly becomes clear that items 11, 25 and 26 are serious off-line outliers regardless of dataset or method of calculating the working memory score. When items 11, 25, and 26 were removed, the correlation between unweighted working memory scores and posterior means for the μ_c parameters increased to .63($p < .001$) for the main dataset and .56 ($p < .01$) for the cross-validation dataset. The correlation between weighted working memory scores and posterior means for the μ_c parameters increased to .60 ($p < .01$) for the main dataset and .55 ($p < .01$) for the cross-validation dataset.

We are left, however, with the three items not behaving as expected. Two of these, items 25 and 26, involved multiple instances of a single rule. Perhaps the working memory demands of $3n$ are less than those of $n + m + o$. However, item 11 had the highest compensation parameter in the entire set of matrices, but still had a low working memory score. Furthermore, the rules involved were quite simple: two instances of the identity rule.

Perhaps analyses concerning the relationship between the presence of particular rules and the compensation parameter could help clarify why some items could have high compensation parameters but low working memory demands. To this end, a set of simple linear multiple regressions was performed on the nonreduced versions of both datasets. Unweighted working memory load and the presence/absence of the most frequently appearing matrix solution rules (identity,quantitative progression, addition/subtraction and rule-of-three) were the predictor variables and $\hat{\mu}_c$ the outcome variable. The presence of the identity rule predicted larger compensation parameters (more noncompensatory processing), but only in the main dataset ($B = .1622(.076)$, $t(21) = 2.12, p = 0.046$). This finding is unexpected. The identity rule is considered the easiest rule to apply; it should tax working memory the least of all rules. It is also likely to be the easiest rule to abstract. Perhaps there is an unmodeled attribute that is confounded with identity and compensation.

As a final note, the correlation between $\hat{\bar{b}}$ and $\hat{\mu}_c$ became substantial, approximately .60, for both main and cross-validation datasets after the three outliers were removed.

Issues in the Application Study

Convergence problems abounded in the application study, most notably for the ability estimates for the compensatory model. This is especially surprising, because the compensatory model is the ”easy” model in the group. Until recently, it was the only MIRT model with estimable parameters. Also, the cross-validation dataset seemed to foster more convergence problems than did the main dataset. It is unclear why this is the case.

The correlation between difficulty and compensation parameters in the reduced datasets was unmodeled in the WINBUGS syntax. This may have slowed mixing of the chain (Gill, 2002). It is unclear what effect, if any, the correlation may have had on parameter recovery.

The positive relationship between working memory load and noncompensatory processing holds only if we remove three outliers from the item pool. This removal raises concerns about forcing the data to fit the theory. Scatterplots are notorious in permitting viewers to see what they wish to see (Rosenthal, 1966). The only consolation in the matter is that two out of the three removed items were completely unlike the remaining items in the nature of their demands on working memory.

CHAPTER VI

SUMMARY AND CONCLUSIONS

The current investigation involved a simulation study assessing parameter recovery in MCMC estimation of the parameters of the GMIRT Rasch model and an application study assessing the degree of noncompensatory processing in a matrix completion task. Additionally, the relationship between the degree of noncompensatory processing and the working memory load of individual items was assessed. What follows is a summary of the findings for each of these points.

Summary

Parameter Recovery for GMIRT Rasch model

Recovery of item-level compensation parameters was adequate under varying conditions of sample size and interability correlation. Recovery of difficulty parameters was quite good when difficulty parameters were constrained equal across dimensions. Additionally, compensation estimates remained adequate when these constraints were added. Recovery of ability parameters, on the other hand, was inadequate in all conditions. There did not appear to be strong effects for sample size or interability correlation for any parameter.

Degree of Noncompensatory Processing in Matrix Completion Task

The original design of the application study called for a comparison of person- and item-fit for the GMIRT, compensatory, and noncompensatory models as the method for assessing the degree of noncompensatory processing in the matrix completion task. These comparisons were not possible given the poor quality of the ability estimates. Descriptive assessments of compensation estimates indicate that a few items were very nearly compensatory, but most items, even with a median compensation parameter as low as .15, exhibited probability contours revealing substantial differences at extremes of ability. These differences in probability of correct response largely affected examinees who are offline outliers as predicted in earlier research (Ackerman, n.d.), but also affected online outliers for exceptionally easy or exceptionally difficult items.

Relationship between working memory load and noncompensatory processing

When the handful of strong offline outliers was removed from the item-pool, working memory load and noncompensatory processing showed the expected positive relationship. The effect was present in both main and cross-validation datasets with both weighted and unweighted working memory scores.

Implications of Results

Study-specific limitations and findings are discussed in the results and discussion section for the particular study. Over-arching themes are discussed here.

Estimation of Noncompensatory Models

The results suggest that the compensation parameter in the GMIRT Rasch model is identifiable, at least when a strong prior is used. It appears, however, that ability parameters in multidimensional models are exceptionally difficult to estimate via MCMC. Finally, despite successful application of the GMIRT model to a task of interest in educational measurement, MCMC estimation for the GMIRT model remains a laborious task. Unless one has a specific research need for the estimates, one should probably use a simpler model. Also, it is entirely clear that the GMIRT model is not ready for use in high stakes testing. In line with past research comparing multiplicative and additive models, the GMIRT model and the noncompensatory model gave very similar test level results.

Noncompensatory Processing in Inductive Reasoning Tasks

In light of the differences among the probability contour plots for models in this study, there is concern whether the compensatory model adequately captures performance for the ”strangely gifted”, those with high ability in one skill and low ability in another, or in the case of very easy items those of uniformly low abilities, or in the case of very difficult items those of very high abilities. Unfortunately, in these last two cases, the models differ just at those ability locations where we need the item to be functioning at its very best.

Working memory and noncompensatory processing

The research on the relationship between working memory and compensation was not neat and clean to begin with. It remains somewhat unkempt after the current study as a handful of items did not show the positive relationship between noncompensatory processing and working memory load.

Directions for Future Research

Future projects stemming naturally from the current studies fall into one of three categories: *fix it*, direct repairs to the current investigation; *follow-up*, natural follow-ups; and *further on*, further investigations of the phenomena of noncompensatory processing.

In the *fix it* realm, several additions and corrections to the current study need to be undertaken before asking reviewers or a wider audience to read the results. First, the simulation study will need to be redone with 50,000 iterations and a second set of estimations on a completely new set of items. The extra iterations may reveal differences in the accuracy of parameter recovery across different sample sizes and interability correlations. Secondly, estimation runs for the BAS matrices that failed convergence should be re-run until a converged run is obtained. Thirdly, item-fit statistics need to be calculated for the GMIRT, noncompensatory and compensatory item-parameter estimates from the application study. Fortunately, Orlando and Thissen's (2001) $S - G^2$ and $S - \chi^2$ statistics do not require ability estimates, but only total scores. Additionally, these statistics have been adapted for use with the multidimensional compensatory model (B. Zhang, 2003). It should not be an overwhelming task to adapt them for use with the GMIRT and noncompensatory models.

Several proximal follow-up studies also come to mind. First, there was a hint in the simulation results that extreme difficulty parameters were associated with less good recovery of the compensation parameters. The relationship of the value of the difficulty parameters and the recovery of the compensation parameter can be systematically manipulated and then examined. This study is planned for the coming year. Secondly, in the application study, compensation and difficulty parameters showed substantial correlation after the three outlying items from the working memory analyses were removed from the data. Further research should be done on the effects of such unmodeled correlation on parameter recovery. Furthermore, potential problems involving the correlation could be avoided altogether by sampling the difficulty and compensation parameters within the same unit in the Gibbs sampler (Gill, 2002; Patz & Junker, 1999b). Additionally, pilot data using the two-dimensional compensatory model, suggested that the final items of the BAS matrices loaded strongly on a second dimension. A freely estimated discrimination parameter would be a welcome addition to the GMIRT model. To date, however, researchers have been able to freely estimate either the discrimination parameters or the compensation parameters at the item level, but not both (Ackerman & Turner, 2003). Finally, estimation of dimension-specific abilities was problematic in the current investigation. This is clearly, a severe problem for estimation in multidimensional models and one that merits researchers' attention. If this problem were solved then true cross-validation would be feasible. Several lines of research have potential to advance the study of compensatory processing to a higher plane. First, many tasks can admit to more than one solution (Mislevy & Verhelst, 1990;Tatsuouka, 1995). Different individuals may be more inclined to use their mental resources compensatively than other individuals. Researchers should develop mixture models for assessing such differences in compensation. If the current estimation problems of the GMIRT model can be solved, perhaps a compensation parameter with random variation at the person-level may be included in the model. A first application of such a model could be to examine the relationship between expertise and compensation. Expertise can offset degradations in performance attributable to increased working memory load (Engle, Kane, & Tuholski, 1999). Finally, IRT modeling might not capture the detail of what one individual is doing when he or she compensates for skill deficits. Talk-aloud protocols (Newell & Simon, 1972) may allow insight into the mental processes a *compensator* uses. For instance, one might administer a handful of the Ravens matrices to ADHD college students, ask them to verbalize their problem-solving experience and thereby gain insight into how they engage the abstraction and working memory demands of the task.

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *Journal of Experimental Psychology: General*, 117:288–318.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory items. *Applied Psychological Measurement*, 13:113–127.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7:255–278.

Ackerman, T. A. (1996). Graphical representation of multidmensional item response theory analyses. *Applied Psychological Measurement*, 20:311–329.

Ackerman, T. A. (2000). Practical applications of item response theory: A short course for measurement, testing and research professionals.

Ackerman, T. A. and Bolt, D. B. (1995). How different cognitive strategies produce differential item performance.

Ackerman, T. A., Neustel, S., and Hombo, C. (2002). Evaluating indices used to assess the goodness-of-fit of compensatory multidimensional item response model.

Ackerman, T. A., Neustel, S., Kelkar, V., and Simpson, M. A. (2001a). An evaluation of the confirmatory estimation procedures of the computer program NOHARM.

Ackerman, T. A., Neustel, S., Kelkar, V., and Simpson, M. A. (2001b). A simulation study examining NOHARM's ability to recover two-dimensional generated item parameters.

Ackerman, T. A. and Turner, R. C. (2003). Estimation and application of a generalized MIRT model: Assessing the degree of compensation between two latent abilities.

Adams, R. J., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21:1–23.

Baddeley, A. and Logie, R. (1999). *Working memory: The multiple-component model*, pages 28–61. Cambridge: Cambridge University Press.

Baeckman, L. and Dixon, R. A. (1992). Psychological compensation: A theoretical framework. *Psychological Review*, 112:259–283.

Baker, F. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker, Inc.

Béguin, A. and Glas, C. E. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Journal of Educational and Behavioral Statistics*, 66:541–562.

Billings, R. S. and Marcus, S. S. (1983). Measures of compensatory and noncompensatory models of decision behavior: processing tracing versus policy capturing. *Organizational Behavior and Human Performance*, 31:331–352.

Bloxom, B. and Vale, C. D. (1987). Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersions of theta.

Bock, R. D., Gibbons, R. D., and Muracki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 12:261–280.

Bolt, D. M. and Lall, T. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, 27:395–414.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–435.

Carlin, B. P. and Louis, T. A. (2001). *Bayes and empirical bayes methods for data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Carpenter, P., Just, M., and Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, 97:404–431.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect model. In Resnick, L. B., editor, *The nature of intelligence*, pages 26–56. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press.

Case, R., Kurland, D. M., and Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, 33:386–404.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46:167–174.

Cattell, R. B. (1940). A culture free intelligence test. *Journal of Educational Psychology*, 31:161–179.

Centre for Longitudinal Studies. University of London . British cohort study of 1970, first followup- 1980 [Data File].

Chi, M. T. H. (1978). Knowledge structures and memory development. In Siegler, R. S., editor, *Children's thinking: what develops*, pages 73–96. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40:5–32.

Cohen, A. S. and Bolt, D. M. (2002). A mixture model analysis of differential item functioning.

Cohen, A. S., Wollack, J. A., Bolt, D. M., and Mroch, A. A. (2002). A mixture rasch model analysis of test speededness.

Congdon, P. (2003). *Applied bayesian modelling*. Chichester, West Sussex, UK: John Wiley & Sons.

Conway, M. and Giannopoulos, C. (1993). Dysphoria and decision making: Limited information use for evaluations of multi-attribute targets. *Journal of Personality and Social Psychology*, 64:613–623.

Coombs, C. (1964). *A theory of data*. New York: John Wiley.

Coombs, C. and Kao, C. (1955). Nonmetric factor analysis. *Engineering Research Bulletin*, 38.

Copi, I. M. (1953). *Introduction to logic*. New York: MacMillan.

Cowen, N. (1999). *An Embedded processes model of working memory*, pages 62–101. Cambridge: Cambridge University Press.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904.

de la Torre, J. and Douglas, J. (2004). Model evaluation and selection in cognitive diagnosis: An analysis of fraction subtraction data.

DeChamplain, A. F. (1999). An overview of nonlinear factor analysis and its relationship to item response theory. Newton, PA: Law School Admission Council. Law School Admission Council Statistical Report 95-3.

DiBello, L. (2002). Skills-based scoring models for the PSAT/NMSQT.

DiBello, L., Stout, W., and Roussos, L. (1995). Cognitively diagnostic assessment. pages 391–410. Hillsdale, NJ: Lawrence Erlbaum Associates.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38:67–86.

Ebbinghaus, H. (2003). Memory: A contribution to experimental psychology. Original work published 1888. Translated by H. A. Ruger & C. E. Bussenius, 1913.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Journal of Personality and Social Psychology*, 64:613–623.

Elliott, C. (1978). *British Ability Scales*. Windsor, UK: National Foundation for Educational Research.

Elliott, C. (1990). *Differential Ability Scales*. San Antonio, TX: The Psychological Corporation.

Embretson, S. E. (1985). Review of the British Ability Scales. *Ninth mental measurements yearbook*, 2:231–232.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3:380–396.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In Irvine, S. H. and Kyllonen, P. C., editors, *Item generation for test development*, pages 219–250. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Embretson, S. E., Schneider, L. M., and Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, 23:13–32.

Engle, R. W., Kane, M. J., and Tuholski, S. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In Miyake, A. and Shah, P., editors, *Models of working memory*, pages 102–134. Cambridge: Cambridge University Press.

Flavell, J. H., Miller, P. H., and Miller, S. A. (1993). *Cognitive development*. Englewood Cliffs, NJ: Prentice Hall, 3 edition.

Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software and manual]. Unpublished manuscript, University of New England, Armidale, Australia.

Fraser, C. and McDonald, R. P. (2003). NOHARM: A DOS computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software and manual]. Retrieved October 10, 2003 from University of Niagara: http://www.niagarac.on.ca/ cfraser/download/nhman.html.

Frye, A. (2001). *An examination of effects on decision accuracy of changes in exam length, case selection, and scoring method in complex performance assessments*. PhD thesis, University of North Carolina at Greensboro.

Ganzach, Y. and Czaczkes, B. (1995). On detecing nonlinear noncompensatory judgment strategies: comparison of alternative regression models. *Organizational Behavior and Human Performance*, 61:168–176.

Garlick, D. (2002). Understanding the nature of the general factor of intelligence: the role of individual differences in neural plasticity as an explanatory mechanism. *Psychological Review*, 109:116–136.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A. W. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511.

Gewecke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J., Berger, Dawid, A. P., and Smith, A., editors, *Bayesian Statistics 4*, pages 169–193. Oxford: The Oxford University Press.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44:445–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov chain monte carlo. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo In Practice*, pages 1–19. Boca Raton, FL: Chapman and Hall/CRC.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, 43:169–177.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman and Hall/CRC.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23:483–496.

Goldberg, L. R. (1971). Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance*, 6:458–479.

Gregory, R. (1992). *Psychological testing: History, principles and applications*. Boston: Allyn and Bacon.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Dordrect, the Netherlands: Kluwer-Nijhoff Publishing.

Hammon, K. R. and Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review*, 72:215–224.

Hartz, S., Roussos, L., and Stout, W. (2002). Prime assessment: Skills diagnosis theory and practice. Unpublished Manuscript.

Hasher, L. and Tacks, R. T. (1979). Automatic versus effortful processes in memory. *Journal of Experimental Psychology: General*, 3:356–388.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Henson, R., Stout, W., He, X., and Douglas, J. (2004). Measures of reliability in models for skills diagnosis.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In Irvine, S. H. and Kyllonen, P. C., editors, *Item generation for test development*, pages 159–178. Mahwah, NJ: Lawrence Erlbaum Associates.

Hornke, L. F. and Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, 4:369–380.

Hsu, Y. (1995). Item parameter estimation of a two-dimensional generalized MIRT model. (Doctoral Dissertation, University of Illinois, 1995). *Dissertation Abstracts International, 57*, 1584.

Hunt, E. B. (1974). Quote the raven? nevermore! In Gregg, L. W., editor, *Knowledge and cognition*, pages 129–158. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Hunt, R. R. and Ellis (1993). *Fundamentals of cognitive psychology*. Madison, WI: WCB Brown and Benchmark, 5 edition.

Insightful Corporation (2001). *S-PLUS*. Seattle, WA: Insightful Corporation.

Jacobs, P. I. and Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement*, 32:235–248.

Jiang, H., DiBello, L., and Stout, W. F. (1996). An estimation procedure for the structural parameters of the unified cognitive/IRT model.

Johnson, H. M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology*, 47.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal data modelling*. New York: Springer.

Jones, D. H. and Nediak, M. (2000). Item parameter calibration of LSAT items using MCMC approximation of bayes posterior distributions. Rutcor research report, rr 7-2000, Rutgers University.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively relevant assessment.* Retrieved October 10, 2002 from Carnegie Mellon University, Statistics Department, http://www.stat.cmu.edu/ brian/nrc/cfa/documents/final.pdf. Prepared for the Committee on the Foundations of Assessment, National Research Council.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25:258–272.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, 42:551–560.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kyllonen, P. (2003). Working memory and knowledge are the key factors underlying item difficulty.

Lim, C. (1993). An application of the joint maximum likelihood estimation procedure to a two-dimensional case of Sympson's noncompensatory IRT model. (Doctoral Dissertation, University of Iowa, 1993). *Dissertation Abstracts International, 54*, 2549.

Logie, R. H. (1996). The seven ages of working memory. In Richardson, J. T. E., Engle, R. W., Hasher, L., Logie, R. H., Stoltzfus, E. R., and Zacks, R. T., editors, *Working memory and human cognition*, pages 31–65. Oxford: Oxford University Press.

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

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20:389–404.

Maris, E. (1992). *Psychometric models for psychological processes and structures*. PhD thesis, University of Leuven, Belgium.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, 60:523–547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64:187–212.

Marshalek, B., Lohman, D. F., and Snow, R. R. (1983). Understanding the nature of the general factor of intelligence: the role of individual differences in neural plasticity as an explanatory mechanism. *Intelligence*, 7:107–127.

Martin, A. D. and Quinn, K. M. (2003). *The MCMCPackage*. Retrieved September 10, 2003 from Washington University, Scythe Statistical Library: http://scythe.wustl.edu/mcmcpack.html.

McDonald, R. (1967). Non-linear factor analysis. *Psychometric monographs*, 15.

McDonald, R. (1997). Normal-ogive multidimensional model. In van der Linden, W. and Hambleton, R., editors, *Handbook of modern item-response theory*, pages 257–269. New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

McKinley, R. (1989). *Confirmatory analysis of test structure using multidimensional item response theory*. Princeton, NJ: Educational Testing Service, June, 1989. Research Report RR-89-31.

McKinley, R. L. and Reckase, M. D. (1982). The use of the general Rasch model with multidimensional item response data. Unpublished manuscript, Iowa City, IA: American College Testing.

Meulders, M., De Boeck, P., and Van Mechelen, I. V. (2003). A taxonomy of latent structure assumptions for probability matrix decomposition models. *Psychometrika*, 68:61–77.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In Nichols, P. D., Chipman, S. F., and Brennan, R. L., editors, *Cognitively diagnostic assessment*, pages 43–71. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Senturk, D., Almond, R., DiBello, L., Jenkins, F., Steinberg, L., and Yan, D. (2001). Modelling conditional probabilities in complex educational assessments.

Mislevy, R. J., Steinberg, L. S., and Almond, R. (In Press). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. (2000). The unity and diversity of exectuive functions and their contributions to complex ”frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41:49–100.

Mulholland, T. M., Pellegrino, J. W., and Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12:252–284.

National Center for Education Statistics and Boston College (2001). TIMSS 1999 mathematics items: Released set for eighth grade.

National Institute of Standards and Technology (2003). NIST/SEMATECH e-handbook of statistical methods. Retrieved October 14, 2003 from http://www.itl.nist.gov/div898/handbook.

Newell, A. and Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Oglivie, J. R. and Schmitt, N. (1979). Situational influences on linear and nonlinear use of information. *Organizational Behavior and Human Performance*, 24:292–306.

Olshavsky, R. W. (1979). Task complexity and contingent processing in decision making: a replication and extension. *Organizational Behavior and Human Performance*, 24:300–316.

Paik, H. S. (1998). *One intelligence or many: Alternative approaches to cognitive abilities*. From http://www.personalityresearch.org/papers/paik.html.

Patz, R. J. and Junker, B. W. (1999a). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24:146–178.

Patz, R. J. and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24:342–366.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, 16:366–387.

Pellegrino, J. W. (1985). Inductive reasoning ability. In Sternberg, R., editor, *Human abilities: An information processing apporoach*, pages 195–225. New York: W. H. Freeman and Company.

Pellegrino, J. W. and Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence*, 3:187–214.

Pelligrino, J. W. and Hunt, E. B. (1991). Cognitive models for understanding and assessing spatial abilities. In Rowe, H. A. H., editor, *Intelligence: Reconceptualization and Measurement*, pages 203–225. Hillsdale, NJ: Lawrence Erlbaum Associates.

Raven, J. C. (1962). *Advanced progressive matrices, set II.* London: H. K. Lewis.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9:401–412.

Reckase, M. D. and McKinley, R. (1991). The discriminating power of test items that measure more than one ability. *Applied Psychological Measurement*, 15:361–373.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-date fit in item response theory. *Applied Psychological Measurement*, 14:127–137.

Richardson, J. T. E. (1996). Evolving concepts of working memory. In Richardson, J. T. E., Engle, R. W., Hasher, L., Logie, R. H., Stoltzfus, E. R., and Zacks, R. T., editors, *Working memory and human cognition*, pages 3–30. Oxford: Oxford University Press.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* New York: Springer.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 45–57. Boca Raton, FL: Chapman and Hall/CRC.

Roesnthal, R. (1966). *Experimenter Effects in Behavioral Research.* New York: Appleton-Century-Crofts.

Salthouse, T. A. (1991). Expertise as the circumvention of human processing limitations. In Ericsson, K. A. and Smith, J., editors, *Toward a general theory of expertise: Prospects and limits*, pages 286–300. Cambridge: Cambridge University Press.

Schneider, W. and Pressley, M. (1997). *Memory development between two and twenty*. Mahwah, NJ: Lawrence Erlbaum Associates, 2 edition.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*, 84:1–66.

Seagull, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61:331–354.

Segall, D. O. (2002). Confirmatory item factor analysis using markov chain monte carlo estimation with applications to online calibration in cat.

Shah, P. and Miyake, A. (1999). *Models of working memory: An introduction*, pages 1–27. Cambridge: Cambridge University Press.

Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology*, 53:255–263.

Snow, R. E., Kyllonen, P. C., and Marshalek, B. (1984). The topography of ability and learning correlations. In Sternberg, R. S., editor, *Advances in the psychology of human intelligence*, pages 47–103. Mahwah, NJ: Lawrence Erlbaum Associates.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: MacMillan.

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS, Version 1.4* [Computer software and manual]. MRC Biostatistics Unit, Institute of Public Health, Cambridge.

Spray, J. and Ackerman, T. A. (1986). An analysis of multidimensional item response theory models.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, 34:353–378.

Sympson, J. B. (1978). A model for testing with multidimensional items. In Weiss, D. J., editor, *Proceedings of the 1977 computerized adaptive testing conference*, pages 82–103.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540.

Tatsuouka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In Nichols, P. D., Chipman, S. F., and Brennan, R. L., editors, *Cognitively diagnostic assessment*, pages 327–360. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric monographs*, 1.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1728.

Tierney, L. (1996). Introduction to general state-space markov chain theory. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 59–74. Boca Raton, FL: Chapman and Hall/CRC.

University of Bristol (1980). 1970 British Cohort Study: Ten-year followup. Technical report, Department of Child Health, University of Bristol.

van der Linden, W. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, 20:373–388.

Van Leeuwe, J. F. J. and Roskam, E. E. (1991). The conjunctive item response model: A probabilistic extension of the Coombs and Kao model. *Methodika*, 5(14-32).

Wang, M. (1986). Estimation of ability parameters from response data to items that are precalibrated with a unidimensional model.

Wang, M. (1987). Fitting a unidimensional model to multidimensional item response data.

Ward, J. and Fitzpatrick, T. F. (1973). Characteristics of matrices items. *Perceptual and motor skills*, 36:987–993.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45:479–494.

Whitely, S. E. (1981). Modelling aptitude test validity from cognitive components. *Journal of Educational Psychology*, 72:750–769.

Whitely, S. E. and Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Journal of Educational Measurement*, 5:383–397.

Whitsett, J. E. (1995). *Boolean algebra and its applications*. Mineola, NY: Dover. Orignal work published 1961.

Wiggins, N. and Hoffman, P. J. (1968). Three models of clinical judgment. *Journal of Abnormal Psychology*, 73:70–77.

Wilson, D. T., Wood, R., and Gibbons, R. D. (1998). Testfact: Test scoring, item statistics, and item factor analysis [Computer software and manual]. Chicago: Scientific Software International.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., and Bock, D. (2003). Testfact4. Chicago: Scientific Software International.

Yntema, D. B. and Torgerson, W. S. (1961). Man-computer cooperation in decisions requiring common sense. *IRE Transactions on Human Factors in Electronics*, 2:20–26.

Zhang, J. (2000). Estimating multidimensional item response models with approximate simple structure.

Zhang, J. (2001). Evaluating the performance of ASSEST: A new item parameter estimation program for multidimensional item response theory models.

Zhang, J. and Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64:213–249.

Appendix A. Tables

Table 1: Q-Matrix for a Hypothetical Mathematics Assessment

	Item 1	Item 2	Item 3	Item 4
Basic Arithmetic	1	1	1	1
Knowledge of Geometry	1	0	0	1
Problem Solving Skills	0	1	0	1

Table 2: α-Vector for a Hypothetical Examinee

	Mastery
Basic Arithmetic	1
Knowledge of Geometry	0
Problem Solving Skills	0

Table 3: Carpenter, Just , & Shell's 1990 Taxonomy of Rules for Ravens Advanced Progressive Matrices

Rule Number	Main Name	Other Names
1	Constant in a Row	Identity
2	Quantitative Pairwise Progression	
3	Figure Addition or Subtraction	
4	Distribution of Three Values	Rule of Three, Latin Square, Variation of Gestalts
5	Distribution of Two Values	

Table 4: Experimental Conditions for Simulation Study

N persons	3000		6000	
n items	25	50	25	50
$r(\theta_1, \theta_2)$				
0				
.30				
.60				

Table 5: Rule Composition for BAS Matrices

Item	Rules	n Rules	Weighted Score	p value (N=8704)
1	Identity	1	1	0.995
2	Identity	1	1	0.996
3	Reversal	1	2	0.963
4	Flip Over	1	2	0.949
5	Rule of 3	1	4	0.965
6	Identity; Rule of 3	2	5	0.925
7	Identity; Addition	2	4	0.806
8	Rule of 3	1	4	0.896
9	Identity; Rule of 3	2	5	0.780
10	Quantitative Progression	1	2	0.801
11	Identity; Identity	2	2	0.615
12	Rule of 3; Rule of 3	2	8	0.654
13	Quantitative Progression	1	2	0.605
14	Identity; Identity; Identity	3	3	0.453
15	Progression	1	2	0.650
16	Quantitative Progression; Subtraction	2	5	0.591
17	Quantitative Progression; Rule of 3	2	6	0.559
18	Rule of 3; Rule of 3; Rule of 3;	3	12	0.427
19	Rule of 3; Rule of 3; Rule of 3;	2	8	0.523
20	Rule of 3; Rule of 3	2	8	0.449
21	Subtraction	1	3	0.456
22	Identity; Identity; Rule of 3	3	6	0.396
23	Sequential Overlay	1	3	0.433
24	Identity; Quantitative Progression; Addition	3	6	0.235
25	Rule of 2; on 4 objects	4	20	0.151
26	Rule of 2	4	20	0.273
27	Identity; Rule of 3; Rule of 3	3	9	0.086
28	Rule of 3; Permutation with 1 Repetition; Permutation with 1 Repetition; Identity	4	11	0.061

Table 6: Free GMIRT Model, Gewecke Z Statistic for b_1, All Conditions

	N = 3000			N = 6000		
	$r(\theta_1, \theta_2) = 0$	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	-7.6260*	-7.3612*	0.8622	-16.0361*	5.2563*	0.3193
3	-1.0026	-6.2603*	-0.2091	9.6701*	2.6140*	2.3765*
4	-1.5019	-5.1295*	2.3657*	-18.8143*	5.1234*	-0.1024
5	0.3275	-7.9502*	2.4955*	-37.2372*	7.9066*	5.3448*
6	-2.0152*	3.6473*	1.3543	14.0960*	-0.9465*	0.1360
7	-0.4139	-3.6871*	-2.3877*	-12.7989*	-1.8553	-2.2585*
8	-2.4967*	-9.8817*	2.4930*	-18.6517*	8.0005*	-0.4100
9	-0.1630	-10.6368*	2.5893*	-35.9374*	7.9612*	6.0094*
10	5.1039*	-4.8679*	-1.6389	-0.7574	3.5303*	2.0768*
11	2.1964*	-9.1808*	1.8429	-13.4661*	8.724*	0.7399
12	2.8748*	-5.4242*	1.0218	-22.9259*	8.0496*	-2.1923*
13	-0.1647	-5.5946*	2.7051*	-30.8946*	8.1608*	3.5837*
14	-3.1530*	-4.5013*	0.0278	-20.7898*	5.7967*	-1.9978*
15	-0.6213	-10.4013*	2.4779*	-9.5459*	-5.5811*	-4.1524*
16	3.3387*	-1.7100	0.2616	-4.7842*	7.6856*	0.7290
17	-1.5215	-2.0193*	-0.0477	-21.054*	8.5878*	-2.6805*
18	-1.3819	-2.0715*	1.7622	-4.5084*	-0.6272	-5.6156*
19	-1.4679	-0.9004	0.7610	-26.7914*	4.9774*	-0.9481
20	2.5970*	0.4442	-1.0001	-10.4563*	-2.3764*	1.3317
21	-1.9694*	-5.8697*	3.4664	-15.019*	3.8427*	1.2539 *
22	-4.1492*	-8.7477*	1.6747	3.0827*	-0.5083	1.1049
23	-0.3644	-8.0885*	1.2078	-6.4077*	6.1404*	3.3629*
24	0.7313	4.1365*	3.0732*	-20.5315*	2.5298*	1.3282
25	4.0737*	6.4705*	-1.0973	30.9626*	-3.1997*	-3.7559*

$^*p < .05$

Table 7: Free GMIRT Model, Geweke Z Statistic for b_2, All Conditions

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) =$	0	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	7.5335*	6.6553*	-0.3566	15.7167*	-4.9874*	-0.0139
3	1.3385	6.5667*	0.2763	-8.9932*	-2.7616*	-2.5001*
4	1.6323	5.2210*	-2.2426*	18.5838*	-5.0675*	0.2147
5	1.0337	9.1424*	-1.3154	30.0844*	-7.5134*	-4.4001*
6	1.4344	-3.3152*	-1.5273	-12.2228*	1.0769	-0.1537
7	0.6467	3.5205*	2.2944*	10.0361*	2.0515*	2.2424*
8	2.6100*	10.2264*	-2.0291*	18.9525*	-7.4308*	0.8782
9	0.5629	11.0717*	-1.9616*	32.6669*	-6.9685*	-5.6948*
10	5.1039*	4.8073*	2.8892*	4.9489*	-2.1102*	-1.7413
11	2.1964*	9.2741*	-1.6321	13.313*	-8.4783*	-0.5798
12	2.8748 *	4.9752*	-1.2372	23.7671*	-7.9828*	3.3703*
13	-0.1647	5.7131*	-2.5726*	30.1671*	-7.7735*	-3.3048*
14	-3.153*	4.0595*	-0.0383	21.0060*	-5.3699*	2.1356*
15	-0.6213	10.3555*	-2.3143*	4.2419*	5.1621*	4.0567*
16	3.3387*	1.5297	0.1354	6.8185*	-7.3679*	0.3692
17	-1.5215	1.9399*	-0.1110	20.8488*	-7.949*	2.9798*
18	-1.3819	2.3026*	-1.3596	3.4404*	0.9213	5.4534*
19	-1.4679	-0.1057	-0.9552	18.8532*	-4.6435*	1.5898
20	2.5970*	-0.5457	0.7689	9.5766*	2.8925*	-1.1027
21	-1.9694*	6.1140*	-3.2687*	15.0691*	-3.4419*	-1.235
22	-4.1492*	8.9153*	-1.4232	-3.5028*	0.4996	-0.5176
23	-0.3644	8.1768*	-1.1751	6.7901*	-5.9427*	-3.1518*
24	0.7313	-4.2371*	-3.5466*	18.0935*	-2.2324*	-1.3792
25	4.0737*	-6.1029*	0.2625	-26.146*	3.6725*	3.7064*

$*p < .05$

Table 8: Free GMIRT Model, Geweke Z Statistic for μ, All Conditions

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) =$	0	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	0.4848	2.3153*	-1.1867	3.7292*	1.3879	-0.3700
3	-1.1505	0.8336	-1.5146	-2.9802*	1.049	-0.4521
4	-0.0145	-0.4515	-0.7235	6.3266*	1.8529	1.0328
5	0.1602	1.6467	0.3491	7.4351*	2.983*	-3.2124
6	-1.6049	1.7508	-1.2676	-7.5356*	0.5244	2.3129*
7	0.9015	1.2699	-0.4302	6.5458*	-0.3248	1.5534
8	0.2116	0.4218	-1.3209	0.2870	1.5787	-1.1059
9	-2.6874*	1.3022	-2.4324*	3.555*	-0.0469	-3.8034*
10	0.3456	2.8381*	-3.1704*	-7.8603*	-0.8197	-0.9015
11	-0.5712	0.599	-1.4618	1.4553	-0.7402	1.0543
12	-0.9423	3.2608*	-1.0596	1.3040	7.4074*	-2.4623*
13	-0.394	-0.1618	-1.6612	3.5335*	0.6231	-0.8732
14	-0.7128	3.3939*	-0.9412	4.1289*	3.0927*	-3.231
15	-1.2625	3.6506*	-1.4093	7.7578*	-1.0867	0.7741
16	-2.9661*	0.4657	-3.1894*	-4.5610*	2.7444*	-1.2780
17	-0.8349	-0.1697	0.0401	3.7016*	3.7247*	-0.9600
18	-0.8650	-0.1651	-2.261*	2.0486*	-2.6535*	2.3410
19	0.4320	0.8584	1.3009	7.9138*	6.2663*	-0.6370
20	-0.4360	1.7157	-0.2153	2.7192*	-2.1798*	-1.5120
21	-0.9945	-0.7174	-0.4188	2.2331*	-0.2598	-0.1121
22	-0.2532	0.8203	-0.2709	-1.9833*	0.5414	-4.1799*
23	0.0132	0.2356	-0.3754	-4.3807*	1.3063	-0.2214
24	-0.5825	-0.1010	0.1871	5.4566*	-0.3526	1.2195
25	-0.4517	-0.5725	1.7604	-12.3445*	-1.6050	-0.1706

*$p < .05$

Table 9: Free GMIRT Model, Proportion of Items Failing Geweke Criterion by Condition*

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Proportion of items		
b_1	3000	0.5000	0.8750	0.3750
	6000	0.9583	0.8333	0.5417
b_2	3000	0.4583	0.8333	0.3750
	6000	1.0000	0.8750	0.5000
μ_c	3000	0.0833	0.2083	0.1667
	6000	0.8750	0.3333	0.2917

*$p < .05$

Table 10: Free GMIRT Model, Proportion of Persons Failing Geweke Criterion by Condition, θ Parameters ($N = 100$)*

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Proportion of persons		
θ_1	3000	0.9400	0.9700	0.8500
	6000	0.9900	0.9700	0.1800
θ_2	3000	0.9300	0.9800	0.8100
	6000	1.0000	0.9500	0.2000

*$p < .05$

Table 11: Free GMIRT Model, Lag-50 Autocorrelations for b_1, All Conditions

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) = 0$		.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	0.0069	0.0304	0.0888	0.5946	0.3434	0.2839
3	0.0659	0.1952	0.1347	0.3135	0.2892	0.1437
4	0.2242	0.1098	0.0567	0.6897	0.4139	0.3006
5	0.3740	0.4241	0.3645	0.8224	0.7224	0.6040
6	0.2806	0.3115	0.3912	0.5286	0.5124	0.4565
7	0.0281	0.0755	0.1209	0.3383	0.2354	0.2920
8	0.0045	0.1121	0.2401	0.5306	0.5591	0.4355
9	0.2216	0.3173	0.3707	0.8229	0.6623	0.3996
10	0.2385	0.1921	0.3107	0.5155	0.2587	0.3406
11	0.1056	0.1390	0.1449	0.6119	0.4411	0.3171
12	0.1705	0.1760	0.2135	0.6440	0.6010	0.5375
13	0.1698	0.0086	0.0803	0.7740	0.5936	0.4362
14	0.1010	0.0117	0.0682	0.6839	0.4674	0.2964
15	0.0208	0.0957	0.0515	0.3959	0.2421	0.1774
16	0.1921	0.1696	0.1764	0.3647	0.2833	0.2792
17	-0.0624	-0.0444	0.0893	0.6434	0.5853	0.2148
18	0.0242	0.0260	0.2473	0.2206	0.1734	0.2305
19	0.1922	0.1620	0.3218	0.6170	0.6264	0.5099
20	0.1717	0.1456	0.1291	0.4601	0.1849	0.1877
21	0.0121	0.1034	0.1121	0.2946	0.0448	0.0099
22	0.2677	0.3474	0.164	0.3768	0.2494	0.4494
23	-0.0461	0.1537	0.1495	0.3719	0.3085	0.2580
24	0.1064	0.1542	0.3074	0.5221	0.1975	0.3230
25	0.1364	0.2778	0.4653	0.7149	0.4851	0.4530

Table 12: Free GMIRT Model, Lag-50 Autocorrelations for b_2, All Conditions

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) =$	0	.3	.6	0	.3	.6
1	......	......	......	......	......	.....
2	-0.0291	0.0205	0.0138	0.5494	0.3486	0.2862
3	0.0617	0.1884	0.1491	0.3029	0.2786	0.1102
4	0.1334	0.1165	0.0510	0.6614	0.4107	0.3023
5	0.3735	0.4124	0.3486	0.8140	0.6694	0.6009
6	0.2501	0.3079	0.3922	0.4667	0.4611	0.4354
7	0.0266	0.0741	0.1068	0.2758	0.1985	0.2737
8	-0.0137	0.1244	0.2109	0.5151	0.5240	0.4412
9	0.2152	0.3126	0.3467	0.8150	0.6178	0.4013
10	0.2408	0.1946	0.2849	0.4381	0.1245	0.2836
11	0.0879	0.1404	0.1441	0.5922	0.4308	0.3159
12	0.1368	0.1252	0.2129	0.6419	0.5892	0.5504
13	0.0855	0.0008	0.0747	0.7589	0.5678	0.4286
14	0.0322	0.0058	0.0673	0.6496	0.4463	0.3002
15	-0.0598	0.0530	0.0481	0.2425	0.1774	0.1681
16	0.1719	0.1583	0.1232	0.2868	0.1858	0.2937
17	-0.0692	-0.0374	0.0872	0.6283	0.5499	0.2060
18	0.0056	0.0049	0.2596	0.1879	0.1400	0.2092
19	0.1572	0.1548	0.2983	0.5368	0.5739	0.5121
20	0.1578	0.1359	0.123	0.3679	0.1366	0.1668
21	0.0054	0.1055	0.1224	0.2839	0.0314	0.0183
22	0.2478	0.3371	0.1526	0.349	0.2019	0.4445
23	-0.0537	0.1507	0.1481	0.3647	0.2915	0.2560
24	0.1079	0.1448	0.2786	0.4994	0.1640	0.3228
25	0.1218	0.2572	0.4583	0.6763	0.4419	0.4123

Table 13: Free GMIRT Model, Lag-50 Autocorrelations for μ

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) =$	0	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	-0.0437	-0.1418	-0.0681	0.1371	-0.0166	-0.0367
3	-0.0313	0.0043	-0.0169	-0.0180	-0.0203	-0.0829
4	0.0832	-0.0367	-0.0411	0.1343	-0.0222	-0.0110
5	-0.0771	-0.1115	-0.0780	0.3600	0.2418	0.1176
6	-0.1268	-0.0050	-0.0448	0.1874	0.0595	0.1173
7	-0.1811	-0.1202	-0.0354	0.0781	0.0019	-0.0182
8	-0.1580	-0.1446	-0.0413	-0.0455	-0.0854	-0.0721
9	-0.1220	-0.1722	-0.0851	0.2070	0.2047	0.0454
10	-0.0334	-0.0673	0.0322	0.1242	-0.0175	0.0612
11	-0.0936	-0.0902	-0.0609	0.0876	0.0076	-0.0230
12	-0.1315	-0.0570	-0.0281	0.1100	0.1258	0.0985
13	-0.0180	-0.0929	-0.0767	0.0355	0.0406	-0.0156
14	-0.0872	-0.0658	-0.0445	0.1631	0.0113	-0.0412
15	-0.1547	-0.1498	-0.1256	0.3357	-0.0594	-0.0022
16	-0.0874	-0.0548	-0.0282	0.1027	-0.0512	0.0251
17	-0.1165	-0.1041	-0.0512	0.0339	-0.1107	-0.0520
18	-0.1955	-0.0993	-0.1048	-0.0463	-0.0517	-0.0116
19	-0.0843	-0.0816	0.0270	0.2805	0.2634	0.1659
20	-0.1782	-0.1055	-0.0648	0.1578	0.0079	-0.0059
21	0.0264	-0.0037	-0.0008	-0.0846	-0.0995	-0.0895
22	-0.1335	-0.0873	-0.1346	0.1103	-0.0177	0.0729
23	-0.0522	-0.0107	0.0174	-0.0766	-0.0643	-0.0547
24	-0.1360	-0.0730	-0.0139	0.0958	-0.0351	0.0463
25	-0.1672	-0.0234	0.0010	0.3451	0.0361	0.1057

Table 14: Free GMIRT Model, Mean Lag-50 Autocorrelations for GMIRT Item Parameters and τ_θ by Condition

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Mean Autocorrelation		
b_1	3000	0.1253(0.113)	0.1539(0.116)	0.2000(0.121)
	6000	0.5355(0.173)	0.395(0.185)	0.3307(0.138)
b_2	3000	0.0997(0.115)	0.1453(0.115)	0.1876(0.120)
	6000	0.496(0.187)	0.3567(0.189)	0.3225(0.141)
μ_c	3000	-0.0958(0.069)	-0.0789(0.049)	-0.0445(0.044)
	6000	0.1173(0.127)	0.0145(0.101)	0.0141(0.071)
τ_θ	3000	0.9365 (0.027)	0.9371(0.007)	0.9279(0.008)
	6000	0.9534 (0.012)	0.9627(0.012)	0.9598(0.005)

Table 15: Free GMIRT Model, Mean Lag-50 Autocorrelations for θ Parameters ($N = 100$)

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Mean Autocorrelation		
θ_1	3000	0.0924(0.076)	0.1435(0.108)	0.1604(0.115)
	6000	0.1699(0.141)	0.1217(0.102)	0.1272(0.100)
θ_2	3000	0.0960(0.080)	0.1248(0.093)	0.1478(0.103)
	6000	0.1649(0.138)	0.1391(0.119)	0.1325(0.104)

Table 16: Free GMIRT Model, Parameter Recovery for b_1, N = 3000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	-0.1623	......	......	......	......	......	......
2	-0.1316	-0.8104	-0.6788	-1.1370	-1.0054	-1.1440	-1.0124
3	0.4287	1.366	0.9373	1.1660	0.7373	1.127	0.6983
4	-0.908	-0.5586	0.3494	-0.6932	0.2148	-0.7551	0.1529
5	0.9338	-0.1234	-1.0572	-0.6130	-1.5468	-0.6477	-1.5815
6	-0.734	-0.5174	0.2166	-0.4936	0.2404	-0.8903	-0.1563
7	-0.0003	0.3326	0.3329	0.1211	0.1214	0.2769	0.2772
8	0.8566	0.6794	-0.1772	0.4832	-0.3734	0.4759	-0.3807
9	0.2955	-0.0561	-0.3516	-0.4126	-0.7081	-0.4507	-0.7462
10	-0.4805	-0.5343	-0.0538	-0.8377	-0.3572	-0.7362	-0.2557
11	-0.3745	-0.0461	0.3284	-0.2419	0.1326	-0.1739	0.2006
12	0.2316	-0.0457	-0.2773	-0.1914	-0.4230	-0.1084	-0.3400
13	1.2995	0.0987	-1.2008	0.0054	-1.2941	-0.0124	-1.3119
14	0.1654	-0.0457	-0.2111	-0.1341	-0.2995	-0.0600	-0.2254
15	-0.5897	-0.8649	-0.2752	-1.0780	-0.3076	-1.1720	-0.5823
16	-0.7814	-0.8708	-0.0894	-1.0890	-1.0613	-1.0390	-0.2576
17	1.6739	0.7735	-0.9004	0.6126	-0.0624	0.6830	-0.9909
18	0.2301	0.3484	0.1183	0.1677	0.6631	0.0549	-0.1752
19	0.0118	-0.5703	-0.5821	-0.6513	-0.5821	-0.5766	-0.5884
20	0.784	-0.1681	-0.9521	-0.1298	-0.9138	-0.0378	-0.8218
21	0.2584	1.287	1.0286	1.1350	0.8766	1.127	0.8686
22	-0.3827	-0.0277	0.355	-0.5087	-0.1260	-0.4364	-0.0537
23	0.354	0.8311	0.4771	0.6418	0.2878	0.6621	0.3081
24	0.4735	0.0886	-0.3849	0.0166	-0.4569	-0.1586	-0.6321
25	-1.8607	-0.2189	1.6418	-0.0650	1.7957	-0.05819	1.8025

Table 17: Free GMIRT Model, Parameter Recovery for b_1, N = 6000

		$r(\theta_1,\theta_2)=0$		$r(\theta_1,\theta_2)=.3$		$r(\theta_1,\theta_2)=.6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.1623	......	......	......	......	......	......
2	-0.1316	-0.8224	-0.6908	-0.6466	-0.5150	-0.8001	-0.6685
3	0.4287	1.2140	-0.7853	1.0310	0.6023	1.1080	0.6793
4	-0.9080	-0.6950	0.2130	-0.4494	0.0103	-0.4931	0.4149
5	0.9338	-0.3084	-1.2422	0.1915	-0.7423	-0.2472	-1.1810
6	-0.7340	-0.3949	0.3391	0.0581	-0.0548	-0.5614	0.1726
7	0.0003	0.1821	0.1824	0.2733	0.2736	0.2420	0.2423
8	0.8566	0.6575	-0.1991	1.0700	0.2134	0.7269	-0.1297
9	0.2955	-0.2702	-0.5657	0.1827	-0.1128	-0.1379	-0.4334
10	-0.4805	-0.6505	-0.1700	-0.6778	-0.1973	-0.6834	-0.2029
11	-0.3745	-0.1487	0.2258	0.0574	0.4319	-0.1336	0.2409
12	0.2316	-0.1124	-0.3440	0.2687	0.0371	-0.0685	-0.3001
13	1.2995	0.0371	-1.2624	0.3831	-0.9164	0.2127	-1.0868
14	0.1654	-0.0686	-0.2340	0.1830	0.0176	-0.0347	-0.2001
15	-0.5897	-0.9885	-0.3988	-1.0840	-0.4943	-1.0290	-0.4393
16	-0.7814	-0.8424	-0.0610	-0.8333	-0.0519	-0.9986	-0.2172
17	1.6739	0.6048	-1.0691	1.0880	-0.0586	0.6653	-1.0086
18	0.2301	0.2679	0.0378	0.1960	-0.0341	0.2344	0.0043
19	0.0118	-0.7019	-0.7137	-0.3242	-0.3360	-0.6564	-0.6682
20	0.7840	-0.0829	-0.8669	-0.0299	-0.8139	-0.0364	-0.8204
21	0.2584	1.2270	0.9686	1.3350	1.0766	1.2530	0.9946
22	-0.3827	-0.2384	0.1443	-0.2632	-0.1195	-0.2167	0.1660
23	0.3540	0.7245	0.3705	0.9480	0.5940	0.7352	0.3812
24	0.4735	0.0978	-0.3757	0.1605	-0.3130	0.0981	-0.3754
25	-1.8607	-0.2696	1.5911	-0.4457	1.4150	-0.3312	1.5920

Table 18: Free GMIRT Model, Parameter Recovery for b_2, N = 3000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	1.1071	......	......	......	......	......	......
2	-1.1019	-0.9975	0.1044	-0.8894	0.2125	-0.9080	0.1939
3	2.3544	1.205	-1.1494	1.2670	-1.0874	1.2290	-1.1254
4	0.1235	-0.6742	-0.7977	-0.5460	-0.6695	-0.5299	-0.6534
5	-0.949	-0.3308	0.6182	-0.0893	0.8597	-0.0809	0.8681
6	-0.0548	-0.5302	-0.4754	-0.6348	-0.5800	-0.4897	-0.4349
7	0.8043	0.1225	-0.6818	0.2153	-0.5890	0.0656	-0.7387
8	0.9281	0.5817	-0.3464	0.7621	-0.1660	0.6562	-0.2719
9	-0.3212	-0.2948	0.0264	-0.0335	0.3547	-0.0591	0.2621
10	-0.4727	-0.8175	-0.3448	-0.6425	-0.1698	-0.7488	-0.2761
11	0.5973	-0.1706	-0.7679	-0.0236	-0.6209	-0.0863	-0.6836
12	0.1013	-0.0995	-0.2008	-0.0789	-0.1802	-0.2430	-0.3443
13	-0.5229	0.1412	0.6641	0.1738	0.6967	0.1549	0.6778
14	0.0462	-0.0955	-0.1417	-0.0550	-0.1012	-0.1302	-0.1764
15	-0.8844	-0.8845	-0.0001	-0.9282	-0.0438	-0.8572	-0.0272
16	-0.5277	-1.142	-0.6143	-1.0680	-0.5403	-1.1180	-0.5903
17	0.1139	0.5401	0.4262	0.6173	0.5034	0.4474	0.3335
18	0.6233	0.3571	-0.2662	0.2231	-0.4002	0.1519	-0.4714
19	-1.0596	-0.7273	0.3323	-0.6441	0.4155	-0.8251	0.2345
20	-0.5248	-0.0962	0.4286	-0.1738	0.3510	-0.2198	0.3050
21	2.9751	1.246	-1.7291	1.3930	-1.5821	1.3420	-1.6331
22	0.4234	-0.2083	-0.6317	-0.0598	-0.4832	-0.1129	-0.5363
23	1.6185	0.7681	-0.8504	0.9395	-0.6790	0.8106	-0.8079
24	0.1487	0.0963	-0.0524	0.0283	-0.1204	0.1475	-0.0012
25	1.6378	-0.1256	-1.7634	-0.3684	-2.0062	-0.4508	-2.0886

Table 19: Free GMIRT Model, Parameter Recovery for b_2, N = 6000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	1.1071	......	......	......	......	......	......
2	-1.1019	-0.7162	0.3857	-0.9278	0.1741	-0.8311	0.2708
3	2.3544	1.3670	-0.9874	1.4550	-0.8994	1.3480	-1.0064
4	0.1235	-0.5226	-0.6461	-0.7641	-0.8876	-0.7678	-0.8913
5	-0.3474	0.6016	0.2972	-0.8956	0.0534	-0.4551	0.4939
6	-0.3097	-0.2549	0.5417	-0.0829	-0.0281	-0.3587	-0.3039
7	0.8043	-0.2757	-0.5286	0.1304	-0.6739	0.1476	-0.6567
8	0.9281	0.7700	-0.1581	0.1967	-0.7314	0.5697	-0.3584
9	-0.3212	-0.2910	-0.0320	-0.7043	-0.3831	-0.2793	0.0419
10	-0.4727	-0.6096	-0.1369	-0.6115	-0.1388	-0.6692	-0.1965
11	0.5973	-0.1021	-0.6994	-0.2971	-0.8944	-0.1272	-0.7245
12	0.1013	-0.0897	-0.1910	-0.5319	-0.6332	-0.1689	-0.2702
13	-0.5229	0.1730	0.6959	-0.2099	0.3130	-0.0073	0.5156
14	0.0462	-0.0818	-0.1280	-0.3745	-0.4207	-0.1249	-0.1711
15	-0.8844	-0.9304	-0.0460	-0.8077	0.0767	-0.8870	-0.0026
16	-0.5277	-0.8427	-0.3150	-0.9305	-0.4028	-0.8324	-0.3047
17	0.1139	0.7214	0.6075	0.1418	0.0279	0.6264	0.5125
18	0.6233	0.3497	-0.2736	0.4046	-0.2187	0.2550	-0.3683
19	-1.0596	-0.6809	0.3787	-1.2200	-0.1604	-0.7324	0.3272
20	-0.5248	-0.0541	0.4707	-0.0073	0.5175	-0.0602	0.4646
21	2.9751	1.3570	0.2433	1.1480	-1.8271	1.1820	-1.7931
22	0.4234	-0.1302	-0.5536	-0.1370	-0.5608	-0.2807	-0.7041
23	1.6185	0.7782	-0.8403	0.4830	-1.1350	0.7024	-0.9161
24	0.1487	0.1183	-0.0304	0.0887	-0.0600	0.0676	-0.0811
25	1.6378	-0.1291	-0.1291	0.0853	-1.5525	-0.1983	-1.8361

Table 20: Free GMIRT Model, Parameter Recovery for μ_c, N = 3000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.2415	......	......	......	......	......	......
2	0.1016	0.1596	0.058	0.1896	0.0880	0.1700	0.0684
3	0.6134	0.3753	-0.2381	0.3915	-0.2219	0.4063	-0.2071
4	0.0261	0.0364	0.0103	0.0255	0.0006	0.0291	0.0030
5	0.2972	0.3751	0.0779	0.4007	0.1035	0.3929	0.0957
6	0.5417	0.3927	-0.149	0.3970	-0.1447	0.4124	-0.1293
7	0.327	0.2712	-0.0558	0.2907	-0.0363	0.2619	-0.0651
8	0.2022	0.3182	0.116	0.2552	0.0530	0.2636	0.0614
9	0.2495	0.2098	-0.0397	0.1746	-0.0749	0.1857	-0.0638
10	0.3721	0.3191	-0.053	0.3176	-0.0545	0.2764	-0.0957
11	0.0353	0.0596	0.0243	0.0523	0.0170	0.0417	0.0064
12	0.276	0.2988	0.0228	0.2954	0.0194	0.2887	0.0127
13	0.0216	0.0983	0.0767	0.0828	0.0612	0.0771	0.0555
14	0.1553	0.1563	0.001	0.1356	-0.0197	0.0949	-0.0604
15	0.1547	0.1044	-0.0503	0.1411	-0.0136	0.1207	-0.0340
16	0.2113	0.3051	0.0938	0.3307	0.1194	0.2963	0.0850
17	0.2758	0.2437	-0.0321	0.2555	-0.0203	0.2601	-0.0157
18	0.3481	0.2011	-0.147	0.3063	-0.0418	0.3263	-0.0245
19	0.3631	0.3076	-0.0555	0.2832	-0.0799	0.2874	-0.0757
20	0.2723	0.3385	0.0662	0.3169	0.0446	0.2753	0.0030
21	0.0534	0.2967	0.2433	0.1948	0.1414	0.1880	0.1346
22	0.2035	0.148	-0.0555	0.2232	0.0197	0.1767	-0.0268
23	0.1134	0.1204	0.007	0.0628	-0.0506	0.0662	0.0472
24	0.3291	0.3151	-0.014	0.3246	-0.0045	0.3098	-0.0193
25	0.2111	0.3309	0.1198	0.3393	0.1282	0.3236	0.1125

Table 21: Free GMIRT Model, Parameter Recovery for μ_c, N = 6000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.2415	......	......	......	......	......	......
2	0.1016	0.0779	-0.0237	0.0681	-0.0336	0.0692	-0.0324
3	0.6134	0.3826	-0.2308	0.3934	-0.2200	0.4089	-0.2045
4	0.0261	0.0246	-0.0015	0.0175	-0.0086	0.0167	-0.0094
5	0.2972	0.3495	0.0523	0.3639	0.0667	0.3712	0.0740
6	0.5417	0.2759	-0.2658	0.2453	-0.2964	0.2823	-0.2594
7	0.327	0.2975	-0.0295	0.2780	-0.0490	0.2573	-0.0697
8	0.2022	0.1880	-0.0142	0.2115	0.0093	0.1864	-0.1760
9	0.2495	0.2226	-0.0269	0.2021	-0.0474	0.1708	-0.0787
10	0.3721	0.2955	-0.0766	0.2819	-0.0902	0.2628	-0.1093
11	0.0353	0.0622	0.0269	0.0348	-0.0005	0.0372	0.0019
12	0.276	0.2994	0.0234	0.2774	0.0014	0.2426	-0.0334
13	0.0216	0.0852	0.0636	0.0843	0.0627	0.0659	0.0443
14	0.1553	0.1312	-0.0241	0.1292	-0.0261	0.1082	-0.0471
15	0.1547	0.1612	0.0065	0.1332	-0.0215	0.1214	-0.0333
16	0.2113	0.2290	0.0116	0.2165	0.0052	0.2072	-0.0041
17	0.2758	0.2532	-0.0226	0.2482	-0.0276	0.2231	-0.0527
18	0.3481	0.2506	-0.0975	0.2127	-0.1354	0.2342	-0.1139
19	0.3631	0.3196	-0.0435	0.3031	-0.0600	0.2577	-0.1054
20	0.2723	0.3043	0.0320	0.2195	-0.0528	0.2188	-0.0535
21	0.0534	0.2031	0.1497	0.2261	0.1727	0.1996	0.1462
22	0.2035	0.2147	0.0112	0.1874	-0.0161	0.1906	-0.0129
23	0.1134	0.1393	0.0259	0.1191	0.0057	0.1182	0.0048
24	0.3291	0.2922	-0.0369	0.2598	-0.0693	0.2583	-0.0708
25	0.2111	0.3276	0.1165	0.3225	0.1140	0.3734	0.1638

Table 22: Free GMIRT Model, Average Signed Bias and RMSE for Estimates of GMIRT Item Parameters

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Parm.	N	Bias	RMSE	Bias	RMSE	Bias	RMSE
b_1	3000	-0.0586(.689)	0.6762	-0.2367(.679)	0.7592	-0.2418(.654)	0.7576
	6000	-0.1390(.690)	0.6884	0.0077(.574)	0.5622	-0.1211(.654)	0.6519
b_2	3000	-0.3422(.650)	0.7222	-0.2761(.650)	0.7198	-0.3304(.653)	0.7448
	6000	-0.2502(.642)	0.6765	-0.4352(.574)	0.7109	-0.3317(.653)	0.7203
μ_c	3000	0.0011(.101)	0.0989	0.0013(.086)	0.0843	-0.0094(.090)	0.0792
	6000	-0.0156(.090)	0.0894	-0.0300(.096)	0.0985	-0.0364(.092)	0.0976

Table 23: Bolt & Lall (2003): RMSEs for b Parameters, Noncompensatory Rasch Model, 25 Items, $N = 3,000$

	$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter		RMSE	
b_1	0.27	0.36	1.16
b_2	0.27	0.35	0.84

Table 24: Free GMIRT Model, Correlations Between True and Estimated Item Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N			
$r(b_1, \hat{b_1})$	3000	.53	.47	.48
$r(b_1, \hat{b_1})$	6000	.52	.69	.58
$r(b_2, \hat{b_2})$	3000	.82	.78	.79
$r(b_2, \hat{b_2})$	6000	.83	.88	.82
$r(\mu_c, \hat{\mu_c})$	3000	.74	.82	.85
$r(\mu_c, \hat{\mu_c})$	6000	.82	.78	.80
$r(b_1, \hat{b_2})$	3000	.50	.55	.54
$r(b_1, \hat{b_2})$	6000	.49	.24	.49
$r(b_2, \hat{b_1})$	3000	.80	.83	.81
$r(b_2, \hat{b_1})$	6000	.81	.65	.77

Table 25: Free GMIRT Model, Mean Posterior Standard Deviation (PSD) for Item Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	PSD	PSD	PSD
b_1	3000	0.8125(.106)	0.8428(.137)	0.8463(.132)
	6000	0.8348(.182)	0.7393(.133)	0.7498(.100)
b_2	3000	0.8376(.121)	0.7867(.119)	0.8122(.121)
	6000	0.8190(.186)	0.7885(.145)	0.7607(.100)
μ_c	3000	0.0970(.035)	0.0932(.036)	0.0891(.037)
	6000	0.0906(.035)	0.0783(.033)	0.0702(.033)

Table 26: Free GMIRT Model, Mean Proportion Reduction (MPR) from Prior Standard Deviation to Posterior Standard Deviation for Item Parameters

		$r(\theta_1,\theta_2)=0$	$r(\theta_1,\theta_2)=.3$	$r(\theta_1,\theta_2)=.6$
Parameter	N	MPR	MPR	MPR
b_1	3000	0.4436(.075)	0.4040(.097)	0.4016(.093)
	6000	0.4097(.129)	0.477(.094)	0.4698(.070)
b_2	3000	0.4077(.086)	0.4437(.084)	0.4257(.086)
	6000	0.4208(.132)	0.4425(.103)	0.4621(.071)
μ_c	3000	0.3928(.219)	0.4161(.223)	0.4417(.230)
	6000	0.4322(.217)	0.509(.210)	0.5604(.206)

Table 27: Free GMIRT Model, Average Signed Bias and RMSE For Estimates of GMIRT Ability Parameters

		$r(\theta_1,\theta_2)=0$		$r(\theta_1,\theta_2)=.3$		$r(\theta_1,\theta_2)=.6$	
Parm.	N	Bias	RMSE	Bias	RMSE	Bias	RMSE
θ_1	3000	-0.3075(.734)	0.7920	-0.2948(.674)	0.7328	-0.3092(.539)	0.6189
	6000	-0.1156(.744)	0.7488	-0.1358(.615)	0.6264	-0.1527(.654)	0.5196
θ_2	3000	-0.0678(.695)	0.6952	-0.1682(.596)	0.6165	-0.2037(.498)	0.5362
	6000	-0.3425(.791)	0.8588	-0.3700(.657)	0.7515	-0.3592(.550)	0.6550
ENC	3000	1.0748(3.612)	3.6962	1.3766(4.082)	4.2264	1.3228(4.146)	4.2693
	6000	0.6186(3.326)	3.3417	0.2598(3.328)	3.2681	0.5984(3.694)	3.6651

Table 28: Free GMIRT Model, Correlations Between True and Estimated Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N			
$r(\theta_1, \hat{\theta}_1)$	3000	.71	.77	.86
$r(\theta_1, \hat{\theta}_1)$	6000	.72	.80	.88
$r(\theta_2, \hat{\theta}_2)$	3000	.67	.80	.87
$r(\theta_2, \hat{\theta}_2)$	6000	.49	.67	.80
$r(\theta_1, \hat{\theta}_2)$	3000	.71	.77	.86
$r(\theta_1, \hat{\theta}_2)$	6000	.73	.78	.87
$r(\theta_2, \hat{\theta}_1)$	3000	.67	.79	.87
$r(\theta_2, \hat{\theta}_1)$	6000	.50	.67	.81
$r(ENC, \widehat{ENC})$	3000	.93	.93	.95
$r(ENC, \widehat{ENC})$	6000	.90	.93	.94

Table 29: Free GMIRT Model, Mean Posterior Standard Deviation (PSD) for Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	PSD	PSD	PSD
θ_1	3000	0.5379(.053)	0.5483(.068)	0.5902(.088)
	6000	0.7132(.099)	0.7132(.099)	0.6086(.075)
θ_2	3000	0.5237(.099)	0.5804(.074)	0.6089(.094)
	6000	0.7119(.100)	0.7119(.100)	0.6016(.073)

Table 30: Free GMIRT Model, Mean Proportion Reduction (MPR) from Prior Standard Deviation to Posterior Standard Deviation for Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	MPR	MPR	MPR
θ_1	3000	0.4621(.053)	0.4517(.068)	0.4097(.088)
	6000	0.2868(.099)	0.2868(.099)	0.3914(.075)
θ_2	3000	0.4763(.050)	0.4196(.074)	0.3911(.094)
	6000	0.2881(.100)	0.2881(.100)	0.3983(.073)

Table 31: Constrained GMIRT model, Geweke Z Statistic for $\bar{b}$, All Conditions

	N = 3000			N = 6000		
	$r(\theta_1, \theta_2) = 0$	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	6.3217*	2.0979*	-1.8698	-11.8003*	-10.6560*	-3.9104*
3	8.8083*	4.1824*	-1.1097	-17.6926*	-16.0128*	-4.3010*
4	3.0715*	2.9616*	-0.8361	-8.4818*	-7.9968*	-4.0175*
5	7.4327*	2.8534*	-1.9999*	-12.0607*	-10.8298*	-4.7262*
6	7.9811*	1.1111	-0.1549	-12.6056*	-10.3422*	-5.2844*
7	5.9459*	2.9669*	-0.6740	-11.1716*	-10.1106*	-3.1052*
8	6.3725*	0.7301	-1.0682	-7.7757*	-6.1148*	0.2605
9	6.5959*	3.3053*	-2.5519*	-11.7845*	-9.2144*	-2.3986*
10	5.8477*	1.6232	-1.5627	-12.4218*	-10.7012*	-4.2487*
11	5.1439*	1.3851	-1.2587	-7.2379*	-7.2901*	-3.0356*
12	3.8390*	2.2144*	-0.1586	-10.1743*	-10.3863*	-2.2947*
13	4.0391*	2.0067*	-0.1951	-5.6480*	-6.3401*	-1.4873
14	6.6994*	2.5484*	-0.4859	-9.8334*	-11.3558*	-3.8712*
15	4.1796*	1.1720	-1.1857	-12.2184*	-10.7008*	-5.4799*
16	4.4106*	2.3567*	-0.6048	-9.8187*	-10.0761*	-4.5297*
17	7.3944*	2.3496*	0.6718	-10.3750*	-8.2416*	-2.0732*
18	5.8119*	2.8924*	-2.1512*	-10.0071*	-11.0357*	-3.4960*
19	6.0484*	3.1726*	-1.1290	-10.7181*	-9.3695*	-3.0245*
20	5.7599*	1.7416	-1.4707	-12.5611*	-9.3627*	-3.1372*
21	8.9020*	1.3457	-0.3026	-8.7025*	-11.6424*	-2.6163*
22	5.5761*	2.2661*	-0.6068	-10.5129*	-10.1920*	-2.1556*
23	5.4502*	1.4976	-0.6092	-6.9401*	-7.2056*	-2.3269*
24	9.0374*	2.4808*	-1.0075	-11.5320*	-11.1788*	-3.0446*
25	5.3680*	3.1994*	-0.7111	-13.3239*	-12.0862*	-4.3983*

*$p < .05$

Table 32: Constrained GMIRT Model, Gewecke Z Statistic for μ

	N = 3000			N = 6000		
	$r(\theta_1, \theta_2) = 0$	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	-0.3482	0.7412	1.6091	2.0994*	-0.9156	-1.8667
3	-2.3242*	-0.7720	1.3826	3.3577*	-1.4152	-4.9707*
4	3.6279*	-1.2380	0.3563	-0.3254	-3.6076*	-0.4587
5	-0.5845	0.1654	1.5661	-1.5455	-4.6073*	-3.9835*
6	-0.2310	2.7349*	-0.6464	-2.644*	-5.8185*	-2.7956*
7	-0.1046	-0.1936	0.2673	-3.2810*	-4.7083*	-5.2727*
8	-0.4332	1.6373	0.8108	-1.9609	-3.3360*	-5.5132*
9	-1.0301	-1.1913	2.5615*	-1.0333	-2.2349*	-4.3249*
10	0.9690	1.6561	0.9635	-2.8891*	-5.112*	-4.0166*
11	-0.6362	1.2587	1.2732	-1.8700	-1.2886	-1.3794
12	2.4925	0.9169	-0.4836	-3.9355*	-2.6366*	-4.6520*
13	1.2557	0.501	-0.4045	-2.9544*	-2.6967*	-3.3645*
14	-1.3744	0.051	-0.0722	-0.4638	0.8831	-1.6920
15	1.5216	1.6824	0.6680	-4.5400*	-5.6013*	-1.6366
16	1.8592	0.7361	-0.1285	-4.0425*	-6.2104*	-3.7492*
17	-2.5781*	0.6221	-1.0271	1.4349	-1.5864	-4.1168*
18	-0.6391	0.1942	1.9723	-2.8067*	-3.1465*	-4.2664*
19	0.8155	-0.1694	0.4145	-4.0508*	-7.1838*	-5.5268*
20	1.6932	1.9448	0.9772	-3.0786*	-4.9713*	-4.2328*
21	-3.1720*	0.8090	0.6317	-0.0199	2.3234*	-2.8086*
22	-0.8223	0.5116	0.0261	-2.9580*	-4.4081*	-4.7543*
23	-0.5836	-0.0668	0.9004	-1.3051	0.8946	-2.4444*
24	-2.1404*	1.3807	0.4748	-3.0365*	-1.7460*	-4.6728*
25	1.7953	0.3151	-0.0233	-4.2377*	-5.7086*	-4.5337*

$*p < .05$

Table 33: Constrained GMIRT Model, Proportion of Items Failing Geweke Criterion by Condition*

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Proportion of items		
$\bar{b}$	3000	1.0000	0.6667	0.1250
	6000	1.0000	1.0000	0.9167
μ_c	3000	0.2500	0.0417	0.0833
	6000	0.6667	0.7083	0.7917

$*p < .05$

Table 34: Constrained GMIRT Model, Proportion of Persons Failing Geweke Criterion by Condition, θ Parameters ($N = 100$)*

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Proportion of persons		
θ_1	3000	0.9400	0.9700	0.8500
	6000	0.9900	0.9700	0.1800
θ_2	3000	0.9300	0.9800	0.8100
	6000	1.0000	0.9500	0.2000

$*p < .05$

Table 35: Constrained GMIRT Model, Lag-50 Autocorrelations for $\bar{b}$, All Conditions

	N = 3000			N = 6000		
$r(\theta_1, \theta_2) = 0$		.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	0.0922	0.0711	0.0196	0.2626	0.0582	0.0134
3	0.0253	0.0356	0.0201	0.1410	0.0555	0.0084
4	0.0138	0.0229	0.0171	0.1313	0.0314	0.0254
5	0.0670	0.0964	0.0294	0.2518	0.0812	0.0326
6	0.0940	0.0757	0.0310	0.3015	0.0855	0.0505
7	0.0862	0.0834	-0.0008	0.1850	0.0885	0.0225
8	0.0408	0.0305	-0.0043	0.0661	0.0207	0.0023
9	0.0488	0.0598	0.0222	0.2265	0.0670	0.0250
10	0.0599	0.0288	0.0010	0.2884	0.0811	0.0174
11	0.0188	0.0346	0.0089	0.1124	0.0311	0.0088
12	0.0542	0.0649	0.0409	0.1976	0.0966	0.0314
13	0.0116	0.0165	0.0025	0.0644	0.0074	0.0069
14	0.0791	0.0930	-0.0028	0.1931	0.0551	0.0175
15	0.0389	0.0428	0.0324	0.2601	0.0798	0.0289
16	0.0855	0.0918	0.0354	0.2667	0.0755	0.0377
17	0.0629	0.0403	-0.0085	0.1708	0.0351	0.0041
18	0.0578	0.0277	0.0221	0.2321	0.0670	0.0330
19	0.1216	0.0840	0.0069	0.2409	0.0626	0.0307
20	0.0699	0.0835	0.0185	0.2110	0.0500	0.0230
21	0.0288	0.0062	-0.0027	0.0695	0.0299	-0.0058
22	0.0621	0.0539	0.0110	0.2669	0.0855	0.0263
23	0.0190	-0.0053	0.0004	0.0872	0.0223	0.0147
24	0.1079	0.0845	0.0332	0.2225	0.0857	0.0397
25	0.0756	0.0878	0.0360	0.2541	0.0682	0.0338

Table 36: Constrained GMIRT Model, Lag-50 Autocorrelations for μ_c, All Conditions

	N = 3000			N = 6000		
	$r(\theta_1, \theta_2) = 0$	.3	.6	0	.3	.6
1	......	......	......	......	......	......
2	0.0038	-0.0041	0.0032	0.0226	-0.003	-0.0060
3	-0.0051	0.0117	0.016	0.0108	0.0135	0.0144
4	-0.0048	0.0024	0.0152	-0.0003	0.0039	0.0126
5	-0.0205	0.0281	0.0146	0.0199	0.0297	0.0361
6	-0.0043	0.0323	0.0164	0.0315	0.0450	0.0461
7	0.0053	0.0195	-0.0104	0.0106	0.0470	0.0330
8	0.0067	0.0165	-0.0077	0.0155	0.0141	0.0184
9	-0.0042	0.0155	0.0124	0.0086	0.0290	0.0232
10	-0.0032	0.0054	0.0032	0.0193	0.0383	0.0203
11	-0.0184	0.0012	-0.0037	0.0165	0.0111	<0.0000
12	0.0163	0.0177	0.0366	0.0158	0.0411	0.0263
13	0.0046	-0.0015	-0.0017	0.0179	0.0022	0.0060
14	0.0095	0.0141	-0.0177	0.0176	-0.0043	0.0044
15	0.0114	0.0205	0.0240	0.0101	0.0252	0.0065
16	-0.0177	0.0102	0.0202	0.0187	0.0201	0.0301
17	0.0010	0.0059	-0.0086	0.0269	0.0134	0.0093
18	0.0125	0.0016	0.0147	0.0293	0.0211	0.0378
19	0.0197	0.0139	-0.0103	0.0253	0.0251	0.0272
20	-0.0012	0.0148	0.0076	0.0100	0.0289	0.0292
21	0.0018	-0.0031	-0.0085	0.0175	0.0120	-0.0055
22	0.0195	0.0274	0.0020	0.0215	0.0275	0.0227
23	0.0068	-0.0032	0.0097	0.0005	0.0079	0.0152
24	0.0045	0.0249	0.0256	0.0044	0.0225	0.0326
25	-0.0023	0.0218	0.0237	0.0233	0.0336	0.0274

Table 37: Constrained GMIRT Model, Mean Lag-50 Autocorrelations for GMIRT Item Parameters and τ_θ by Condition

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Mean Autocorrelation		
$\bar{b}$	3000	0.0592(0.03)	0.0546(0.031)	0.0154(0.015)
	6000	0.1960(0.074)	0.0592(0.025)	0.0220(0.014)
μ_c	3000	0.0017(0.011)	0.0122(0.011)	0.0074(0.014)
	6000	0.0164(0.008)	0.021(0.014)	0.0195(0.014)
τ_θ	3000	0.9485(0.027)	0.928(0.009)	0.9552(0.003)
	6000	0.9758(0.012)	0.9713(0.012)	0.9649(0.007)

Table 38: Constrained GMIRT Model, Mean Lag-50 Autocorrelations for θ Parameters ($N = 100$)

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	Mean Autocorrelation		
θ_1	3000	0.0659 (0.061)	0.1164 (0.096)	0.1454(0.114)
	6000	0.0368 (0.045)	0.0597(0.058)	0.1191(0.100)
θ_2	3000	0.0734 (0.068)	0.1114 (0.093)	0.1642(0.131)
	6000	0.0432(0.052)	0.0661 (0.064)	0.1240(0.104)

Table 39: Constrained GMIRT Model, Parameter Recovery for $\hat{\hat{b}}$, N = 3000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.4724	......	......	......	......	......	......
2	-0.6168	-0.8751	-0.2584	-0.9530	-0.3362	-0.9616	-0.3448
3	1.3916	1.3650	-0.0266	1.3120	-0.0796	1.2660	-0.1256
4	-0.3922	-0.5965	-0.2042	-0.6061	-0.2138	-0.6288	-0.2366
5	-0.0076	-0.1664	-0.1588	-0.2182	-0.2106	-0.2441	-0.2365
6	-0.3944	-0.4881	-0.0937	-0.5453	-0.1509	-0.5978	-0.2034
7	0.4020	0.2499	-0.1521	0.1965	-0.2055	0.1901	-0.2119
8	0.8924	0.6732	-0.2192	0.6690	-0.2234	0.6263	-0.2660
9	-0.0129	-0.1408	-0.1280	-0.1232	-0.1104	-0.1791	-0.1662
10	-0.4767	-0.6287	-0.1521	-0.6620	-0.1854	-0.6551	-0.1785
11	0.1114	-0.0964	-0.2078	-0.1160	-0.2274	-0.1152	-0.2266
12	0.1664	-0.0240	-0.1904	-0.0814	-0.2479	-0.1331	-0.2996
13	0.3883	0.1400	-0.2483	0.1063	-0.2820	-0.0894	-0.2989
14	0.1059	-0.0518	-0.1576	-0.0772	-0.1830	-0.0823	-0.1881
15	-0.7370	-0.8432	-0.1062	-0.9515	-0.2145	-0.9646	-0.2276
16	-0.6546	-0.9617	-0.3072	-1.0390	-0.3844	-1.0320	-0.3774
17	0.8939	0.6746	-0.2193	0.6380	-0.2559	0.5887	-0.3052
18	0.4267	0.3796	-0.0471	0.2356	-0.1911	0.1879	-0.2388
19	-0.5239	-0.6100	-0.0861	-0.6198	-0.0959	-0.6822	-0.1583
20	0.1296	-0.1217	-0.2513	-0.1316	-0.2612	-0.1215	-0.2511
21	1.6168	1.2970	-0.3198	1.2830	-0.3338	1.2590	-0.3578
22	0.0204	-0.0788	-0.0992	-0.2026	-0.2230	-0.2210	-0.2414
23	0.9862	0.8118	-0.1744	0.7989	-0.1874	0.7471	-0.2392
24	0.3111	0.1114	-0.1997	0.0481	-0.2630	0.0318	-0.2793
25	-0.1114	-0.1501	-0.0386	-0.2182	-0.1068	-0.2683	-0.1568

Table 40: Constrained GMIRT Model, Parameter Recovery for $\hat{\hat{b}}$, N = 6000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.4724	......	......	......	......	......	......
2	-0.6168	-0.7194	-0.1026	-0.7603	-0.1436	-0.7943	-0.1776
3	1.3916	1.3310	-0.0606	1.2730	-0.1186	1.2720	-0.1196
4	-0.3922	-0.5790	-0.1872	-0.5960	-0.2038	-0.6225	-0.2302
5	-0.0076	-0.0866	-0.0790	-0.1621	-0.1545	-0.2533	-0.2457
6	-0.3944	-0.3643	0.0301	-0.4110	-0.0166	-0.4523	-0.0579
7	0.4020	0.2527	-0.1493	0.2207	-0.1813	0.2013	-0.2007
8	0.8924	0.7561	-0.1362	0.7015	-0.1908	0.6853	-0.2070
9	-0.0129	-0.1155	-0.1026	-0.1546	-0.1418	-0.1726	-0.1598
10	-0.4767	-0.5663	-0.0897	-0.6267	-0.1501	-0.6564	-0.1798
11	0.1114	-0.0976	-0.2090	-0.1086	-0.2200	-0.1202	-0.2316
12	0.1664	0.0025	-0.1639	-0.0328	-0.1992	-0.0542	-0.2207
13	0.3883	0.1705	-0.2178	0.1261	-0.2622	0.1196	-0.2687
14	0.1059	-0.0130	-0.1188	-0.0553	-0.1611	-0.0607	-0.1665
15	-0.7370	-0.9114	-0.1744	-0.9493	-0.2122	-0.9575	-0.2204
16	-0.6546	-0.7978	-0.1432	-0.8266	-0.1720	-0.8635	-0.2090
17	0.8939	0.7439	-0.1500	0.6964	-0.1975	0.6730	-0.2209
18	0.4267	0.3218	-0.1049	0.2980	-0.1287	0.2506	-0.1761
19	-0.5239	-0.5616	-0.0377	-0.6089	-0.0850	-0.6213	-0.0974
20	0.1296	-0.0270	-0.1566	-0.0031	-0.1326	-0.0332	-0.1628
21	1.6168	1.3190	-0.2978	1.2590	-0.3577	1.2350	-0.3818
22	0.0204	-0.1840	-0.2044	-0.1914	-0.2117	-0.2095	-0.2299
23	0.9862	0.7661	-0.2202	0.7379	-0.2483	0.7346	-0.2516
24	0.3111	0.1756	-0.1355	0.1482	-0.1629	0.1075	-0.2036
25	-0.1114	-0.2569	-0.1454	-0.2163	-0.1048	-0.2595	-0.1480

Table 41: Constrained GMIRT Model, Parameter Recovery for μ_c, N = 3000

		$r(\theta_1,\theta_2) = 0$		$r(\theta_1,\theta_2) = .3$		$r(\theta_1,\theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.2415	......	......	......	......	......	......
2	0.1016	0.1669	0.0653	0.1932	0.0916	0.1703	0.0687
3	0.6134	0.3904	-0.2230	0.4241	-0.1893	0.4404	-0.1730
4	0.0261	0.0350	0.0089	0.0255	-0.0006	0.0297	0.0036
5	0.2972	0.4182	0.1210	0.4245	0.1273	0.4172	0.1200
6	0.5417	0.4573	-0.0844	0.4748	-0.0669	0.4797	-0.0620
7	0.3270	0.2927	-0.0343	0.3151	-0.0119	0.2969	-0.0301
8	0.2022	0.3342	0.1320	0.2514	0.0492	0.2511	0.0489
9	0.2495	0.2179	-0.0316	0.1710	-0.0785	0.1768	-0.0727
10	0.3721	0.3481	-0.0240	0.3358	-0.0363	0.2908	-0.0813
11	0.0353	0.0588	0.0234	0.0504	0.0151	0.0407	0.0054
12	0.276	0.3061	0.0304	0.3136	0.0376	0.3166	0.0406
13	0.0216	0.0956	0.0739	0.0827	0.0611	0.0761	0.0545
14	0.1553	0.1645	0.0092	0.1408	-0.0145	0.1011	-0.0542
15	0.1547	0.1048	-0.0499	0.1422	-0.0125	0.1242	-0.0305
16	0.2113	0.3057	0.1244	0.3653	0.1540	0.3227	0.1114
17	0.2758	0.2596	-0.1457	0.2693	-0.0065	0.2868	0.0110
18	0.3481	0.2024	-0.0184	0.3239	-0.0242	0.3271	-0.0210
19	0.3631	0.3447	0.1255	0.3167	-0.0464	0.3393	-0.0238
20	0.2723	0.3978	0.2481	0.3607	0.0884	0.3227	0.0504
21	0.0534	0.3015	-0.0624	0.1901	0.1367	0.1782	0.1248
22	0.2035	0.1411	-0.0555	0.2198	0.0163	0.1754	-0.0281
23	0.1134	0.1169	0.0035	0.0606	-0.0528	0.0620	-0.0514
24	0.3291	0.3613	0.0322	0.3727	-0.0436	0.3592	0.0301
25	0.2111	0.3771	0.1660	0.4230	0.2119	0.4432	0.2321

Table 42: Constrained GMIRT Model, Parameter Recovery for μ_c, N = 6000

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Item	Value	Estimate	Bias	Estimate	Bias	Estimate	Bias
1	0.2415	......	......	......	......	......	......
2	0.1016	0.0737	-0.0279	0.0711	-0.0305	0.0737	-0.0279
3	0.6134	0.4650	-0.1484	0.5009	-0.1125	0.4851	-0.1283
4	0.0261	0.0221	-0.0040	0.0175	-0.0086	0.0176	-0.0085
5	0.2972	0.3071	0.0099	0.3426	0.0454	0.4002	0.1030
6	0.5417	0.3455	-0.1962	0.3353	-0.2064	0.3363	-0.2054
7	0.327	0.3184	-0.0086	0.3062	-0.0208	0.3005	-0.0265
8	0.2022	0.1792	-0.0230	0.1987	-0.0035	0.1853	-0.0169
9	0.2495	0.1893	-0.0602	0.1933	-0.0562	0.1773	-0.0722
10	0.3721	0.3012	-0.0709	0.3029	-0.0692	0.2983	-0.0738
11	0.0353	0.0604	0.0251	0.0349	-0.0004	0.0381	0.0028
12	0.276	0.2747	-0.0013	0.2679	-0.0081	0.2480	-0.0280
13	0.0216	0.0703	0.0487	0.0784	0.0568	0.0661	0.0445
14	0.1553	0.1221	-0.0332	0.1258	-0.0295	0.1128	-0.0425
15	0.1547	0.1600	0.0053	0.1555	0.0008	0.1418	-0.0129
16	0.2113	0.2287	0.0174	0.2180	0.0067	0.2156	0.0043
17	0.2758	0.2237	-0.0521	0.2308	-0.0450	0.2329	-0.0429
18	0.3481	0.2735	-0.0746	0.2505	-0.0976	0.2733	-0.0748
19	0.3631	0.2860	-0.0771	0.2809	-0.0822	0.2680	-0.0951
20	0.2723	0.3058	0.0335	0.2347	-0.0376	0.2456	-0.0267
21	0.0534	0.1995	0.1461	0.2335	0.1801	0.2056	0.1522
22	0.2035	0.2485	0.0450	0.2095	0.0060	0.2006	-0.0029
23	0.1134	0.1496	0.0362	0.1163	0.0029	0.1171	0.0037
24	0.3291	0.2754	-0.0537	0.2752	-0.0539	0.2846	-0.0445
25	0.2111	0.5088	0.2977	0.4372	0.2261	0.4592	0.2481

Table 43: Constrained GMIRT model, Average Signed Bias and RMSE for Estimates of GMIRT Item Parameters

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Parm.	N	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\bar{b}$	3000	-0.1686(.080)	0.1860	-0.2155(.076)	0.2279	-0.2423(.065)	0.2506
	6000	-0.1397(.068)	0.1550	-0.1732(.066)	0.1851	-0.1986(.063)	0.2081
μ_c	3000	0.0197(.103)	0.1023	0.0205(.087)	0.0874	0.0114(.084)	0.0833
	6000	-0.0069(.094)	0.0928	-0.0140(.086)	0.0859	-0.0155(.090)	0.0891

Table 44: Constrained GMIRT Model, Correlations Between True and Estimated Item Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N			
$r(b_1, \hat{\bar{b}})$	3000	.99	.99	.99
$r(\mu_c, \hat{\mu_c})$	3000	.74	.82	.80
$r(b_1, \hat{\bar{b}})$	6000	.74	.82	.80
$r(\mu_c, \hat{\mu_c})$	6000	.79	.82	.80

Table 45: Constrained GMIRT Model, Mean Posterior Standard Deviation (PSD) for Item Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	PSD	PSD	PSD
$\hat{\bar{b}}$	3000	0.0806(.019)	0.0757(.018)	0.0718(.016)
	6000	0.0656(.014)	0.0559(.011)	0.0521(.010)
μ_c	3000	0.0366(.032)	0.0761(.029)	0.0706(.028)
	6000	0.0624(.026)	0.0562(.024)	0.0525(.023)

Table 46: Constrained GMIRT Model, Mean Proportion Reduction (MPR) from Prior Standard Deviation to Posterior Standard Deviation for Item Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	MPR	MPR	MPR
$\hat{\bar{b}}$	3000	0.9430(.013)	0.9464(.013)	0.9492(.011)
	6000	0.9536(.010)	0.9605(.008)	0.9631(.007)
μ_c	3000	0.4764(.198)	0.5237(.181)	0.5581(.174)
	6000	0.6094(.165)	0.6480(.154)	0.6711(.144)

Table 47: Constrained GMIRT Model, Average Signed Bias and RMSE for Estimates of GMIRT Ability Parameters

		$r(\theta_1, \theta_2) = 0$		$r(\theta_1, \theta_2) = .3$		$r(\theta_1, \theta_2) = .6$	
Parm.	N	Bias	RMSE	Bias	RMSE	Bias	RMSE
θ_1	3000	-0.3231(.638)	0.7961	-0.3140(.663)	0.7303	-0.3536(.534)	0.6387
	6000	-0.1199(.716)	0.7299	-0.1383(.625)	0.6375	-0.1519(.502)	0.5220
θ_2	3000	-0.0464(.581)	0.6954	-0.1418(.591)	0.6050	-0.1528(.518)	0.5376
	6000	-0.3364(.772)	0.8382	-0.3658(.657)	0.7489	-0.3592(.590)	0.6532
ENC	3000	0.9569(3.543)	3.5976	0.9645(3.973)	4.0071	0.8453(3.892)	3.9024
	6000	0.3635(3.135)	3.0815	0.3635(3.478)	3.4239	0.4387(3.740)	3.6870

Table 48: Constrained GMIRT Model, Correlations Between True and Estimated Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter N				
$r(\theta_1, \hat{\theta_1})$	3000	.71	.77	.86
$r(\theta_1, \hat{\theta_1})$	6000	.73	.80	.88
$r(\theta_2, \hat{\theta_2})$	3000	.67	.80	.87
$r(\theta_2, \hat{\theta_2})$	6000	.50	.67	.80
$r(\theta_1, \hat{\theta_2})$	3000	.71	.77	.86
$r(\theta_1, \hat{\theta_2})$	6000	.72	.80	.88
$r(\theta_2, \hat{\theta_1})$	3000	.67	.79	.87
$r(\theta_2, \hat{\theta_1})$	6000	.49	.67	.81
$r(ENC, \widehat{ENC})$	3000	.93	.93	.95
$r(ENC, \widehat{ENC})$	6000	.90	.93	.94

Table 49: Constrained GMIRT Model, Mean Posterior Standard Deviation (PSD) for Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	PSD	PSD	PSD
θ_1	3000	0.6438(.062)	0.6003(.072)	0.6231(.099)
	6000	0.6796(.071)	0.6368(.058)	0.5809(.072)
θ_2	3000	0.6089(.061)	0.6117(.074)	0.5860(.092)
	6000	0.6383(.070)	0.6152(.055)	0.5692(.071)

Table 50: Constrained GMIRT Model, Mean Proportion Reduction (MPR) from Prior Standard Deviation to Posterior Standard Deviation for Ability Parameters

		$r(\theta_1, \theta_2) = 0$	$r(\theta_1, \theta_2) = .3$	$r(\theta_1, \theta_2) = .6$
Parameter	N	MPR	MPR	MPR
θ_1	3000	0.3562(.062)	0.3997(.071)	0.3769(.099)
	6000	0.3204(.071)	0.3632(.058)	0.4191(.072)
θ_2	3000	0.3911(.061)	0.3883(.074)	0.4140(.092)
	6000	0.3617(.072)	0.3848(.055)	0.4308(.071)

Table 51: Application Study: Proportion of Items or Persons Failing Geweke Statistic by Parameter, Dataset, and Model*

Dataset	A		B	
Chain	1	2	1	2
Model				
GMIRT				
$\bar{b}$	0.0714	0.0357	0.0000	0.6786
μ_c	0.2500	0.5000	0.0714	0.6071
τ_θ	0.7500	1.0000	0.7500	1.0000
θ_1	0.1400	0.2700	0.0400	0.6600
θ_2	0.0500	0.0500	0.0500	0.4500
Noncompensatory				
$\bar{b}$	0.0000	0.0000	0.0000	0.1071
τ_θ	1.0000	1.0000	1.0000	1.0000
θ_1	0.0500	0.0900	0.9300	0.3500
θ_2	0.0000	0.0100	0.9600	0.8000
Compensatory				
d	0.0000	0.0000	0.0000	0.0000
τ_θ	1.0000	1.0000	1.0000	1.0000
θ_1	1.0000	0.9900	0.9500	1.0000
θ_2	1.0000	0.9700	0.9500	1.0000

$*p < .05$

Table 52: Application Study: Mean Autocorrelation by Parameter, Dataset, and Model

Dataset	A		B	
Chain	1	2	1	2
$\bar{b}$	0.0063(0.030)	0.0055(0.032)	0.0727(0.058)	0.1857(0.158)
μ_c	-0.0097(0.025)	-0.0100(0.023)	0.0415(0.091)	0.1259(0.204)
τ_θ	0.5636(0.170)	0.5175(0.167)	0.5593(0.184)	0.9730(0.012)
θ_1	0.0012(0.008)	0.0024(0.008)	0.0007(0.009)	0.0078(0.021)
θ_2	0.0016(0.008)	0.0018(0.008)	0.0007(0.009)	0.0078(0.021)
$\bar{b}$	0.0883(.036)	0.1013(.039)	0.0641(.022)	0.0979(.038)
τ_θ	0.8687(.001)	0.8354(.004)	0.8497(.005)	0.8272(.012)
θ_1	0.0076(.010)	0.0049(.009)	0.0165(.017)	0.0060(.009)
θ_2	0.0131(.011)	0.0102(.010)	0.0134(.013)	0.0110(.011)
d	0.0068(.009)	-0.0004(.012)	-0.0024(.011)	0.0048(.009)
τ_θ	0.9392(.012)	0.9418(.019)	0.9503(.001)	0.9361(.004)
θ_1	0.1833(.0159)	0.1194(.116)	0.1315(.136)	0.1503(.152)
θ_2	0.2086(.179)	0.1272(.124)	0.1277(.132)	0.1529(.154)

Table 53: Multivariate Potential Scale Reduction Factor (MPSRF) for Each Model and Dataset

	Dataset	
Model	A	B
GMIRT	1.0113	......^{a}
Noncompensatory	1.0029	1.0016^b
Compensatory	1.0000^b	1.0002^b

^{a} Only 1 chain available

^{b} θ parameters omitted

Table 54: Posterior Means for Difficulty Parameters, All Models

Model	GMIRT		Noncompensatory		Compensatory	
Dataset	A	B	A	B	A	B
1	.	.	.			
2	-3.9599	-4.0148	-5.1372	-5.4897	5.486	5.721
3	-1.8828	-2.3156	-2.8318	-3.0458	3.1457	3.341
4	-1.8182	-2.1557	-2.5994	-2.5482	2.8902	2.8168
5	-1.7803	-2.0355	-2.9286	-3.0244	3.2525	3.3251
6	-1.1369	-1.3951	-2.025	-2.1581	2.3062	2.4218
7	-0.6042	-0.6591	-0.799	-0.9053	0.9773	1.0798
8	-1.0565	-1.1489	-1.6434	-1.751	1.8911	1.9896
9	-1.0271	-1.3974	-0.6388	-0.7288	0.7794	0.8593
10	-0.4417	-0.4956	-0.7422	-0.8696	0.918	1.0462
11	-0.7489	-0.8001	0.259	0.2104	-0.2406	-0.1938
12	-0.2012	-0.4011	0.113	0.0247	-0.0378	0.0425
13	0.1003	-0.0593	0.3593	0.2711	-0.3126	-0.2317
14	0.487	0.4621	1.0673	0.9814	-1.1262	-1.0396
15	-0.0364	-0.1831	0.1212	0.0603	-0.0433	0.0069
16	0.1084	0.0789	0.4329	0.3325	-0.4001	-0.2954
17	0.0475	-0.0979	0.6303	0.4262	-0.6329	-0.414
18	0.6947	0.6278	1.2328	1.0723	-1.3072	-1.1313
19	0.3498	0.1731	0.7819	0.625	-0.7909	-0.6326
20	0.5233	0.249	1.1061	0.9733	-1.1715	-1.0395
21	0.5808	0.5498	1.0607	0.9638	-1.1164	-1.0178
22	0.8753	0.7684	1.3514	1.2531	-1.4369	-1.3368
23	0.7515	0.7392	1.1563	1.1011	-1.2108	-1.1604
24	1.4473	1.232	2.1615	2.047	-2.4109	-2.3006
25	2.0252	1.9729	2.6794	2.6274	-3.0373	-2.9993
26	1.4961	1.4174	1.9454	1.8437	-2.1514	-2.0524
27	1.904	2.1011	3.2028	3.2433	-3.7472	-3.8006
28	2.3207	2.1997	3.6119	3.4433	-4.2826	-4.0851

Table 55: Posterior Means for Compensation Parameters

Item			Item		
Dataset	A	B		A	B
1	-.—-	-.—-	15	0.1417	0.1846
2	0.1109	0.0676	16	0.1891	0.1525
3	0.0642	0.1088	17	0.3301	0.3036
4	0.0879	0.1661	18	0.1892	0.1456
5	0.0326	0.0593	19	0.1971	0.2263
6	0.0172	0.0518	20	0.2487	0.3705
7	0.1029	0.0868	21	0.1838	0.1499
8	0.0607	0.059	22	0.1231	0.1389
9	0.3677	0.5178	23	0.1125	0.0902
10	0.0563	0.0403	24	0.1479	0.2565
11	0.7476	0.7162	25	0.0106	0.0137
12	0.2286	0.2901	26	0.0107	0.0124
13	0.1612	0.2046	27	0.6069	0.3628
14	0.2532	0.2147	28	0.5103	0.4852

Table 56: Frequency Distribution for μ_c Estimates

μ_c	Dataset A	Dataset B
$\mu_c \leqslant .20$	20	18
$.20 < \mu_c \leqslant .40$	4	7
$\mu_c > .40$	3	3

Table 57: Posterior Standard Deviations, Difficulty Parameters, All Models

Model	GMIRT		Noncompensatory		Compensatory	
Dataset	A	B	A	B	A	B
21	0.0562	0.0594	0.047	0.0468	0.0458	0.045
22	0.0484	0.0561	0.0475	0.0472	0.0467	0.0452
23	0.0471	0.0501	0.0471	0.0471	0.0462	0.0451
24	0.0562	0.0698	0.0491	0.0492	0.0511	0.0503
25	0.046	0.0465	0.0519	0.0527	0.0564	0.0564
26	0.0408	0.0427	0.0481	0.0484	0.0495	0.0485
27	0.0778	0.0868	0.0568	0.0581	0.0665	0.0679
28	0.0895	0.0924	0.0621	0.0612	0.0763	0.0738

Table 58: Proportion Reduction in Standard Deviations for Difficulty Parameters, All Models

Model	GMIRT		Noncompensatory		Compensatory	
Dataset	A	B	A	B	A	B
1	.	.	.			
2	0.6053	0.5743	0.844	0.8322	0.852	0.8726
3	0.9121	0.7917	0.9381	0.9347	0.9382	0.9347
4	0.8819	0.7932	0.9437	0.9444	0.9433	0.9457
5	0.9352	0.8676	0.9361	0.9349	0.936	0.935
6	0.9629	0.936	0.9513	0.9513	0.952	0.9516
7	0.9569	0.9533	0.9621	0.9626	0.9638	0.9644
8	0.9489	0.939	0.9561	0.9557	0.9569	0.957
9	0.9104	0.8834	0.9635	0.9637	0.9646	0.9656
10	0.9629	0.9632	0.9628	0.9629	0.964	0.9645
11	0.9374	0.9282	0.9663	0.9665	0.9674	0.9685
12	0.9517	0.932	0.9658	0.9661	0.9671	0.968
13	0.9599	0.9512	0.9663	0.9665	0.9678	0.9686
14	0.9559	0.9558	0.9668	0.9668	0.9675	0.9684
15	0.9602	0.9509	0.9661	0.9662	0.9673	0.9681
16	0.9567	0.9575	0.9663	0.9667	0.9675	0.9686
17	0.9449	0.9365	0.9667	0.9668	0.9679	0.9687
18	0.9615	0.9567	0.9666	0.9668	0.9673	0.9683
19	0.9584	0.9496	0.9667	0.9667	0.9678	0.9685
20	0.9563	0.9375	0.9667	0.967	0.9674	0.9686
21	0.9603	0.958	0.9668	0.9669	0.9676	0.9682
22	0.9658	0.9604	0.9664	0.9666	0.967	0.968
23	0.9667	0.9646	0.9667	0.9667	0.9673	0.9681
24	0.9603	0.9507	0.9653	0.9652	0.9639	0.9644
25	0.9675	0.9671	0.9633	0.9627	0.9601	0.9601
26	0.9712	0.9698	0.966	0.9658	0.965	0.9657
27	0.945	0.9386	0.9599	0.9589	0.953	0.952
28	0.9367	0.9347	0.9561	0.9567	0.946	0.9478

Table 59: Posterior Standard Deviations for Compensation Parameters

	Item			Item	
Dataset	A	B		A	B
1	-.—-	-.—-	15	0.0273	0.0352
2	0.0861	0.0646	16	0.0356	0.0305
3	0.0225	0.0604	17	0.0569	0.0584
4	0.0375	0.08	18	0.0361	0.034
5	0.0133	0.0292	19	0.0359	0.0427
6	0.0085	0.0162	20	0.0450	0.0684
7	0.0213	0.0195	21	0.0371	0.0368
8	0.0179	0.0182	22	0.0281	0.0318
9	0.0738	0.1029	23	0.0258	0.0247
10	0.0164	0.0132	24	0.0414	0.0653
11	0.0798	0.0854	25	0.0083	0.0146
12	0.0378	0.0523	26	0.0085	0.0158
13	0.0304	0.0391	27	0.1072	0.1125
14	0.0447	0.0411	28	0.1267	0.1353

Table 60: Proportion Reduction in Standard Deviations for Compensation Parameters

Item			Item		
Dataset	A	B		A	B
1	-.—-	-.—-	15	0.8289	0.7793
2	0.4608	0.5955	16	0.7773	0.8091
3	0.859	0.6218	17	0.6438	0.6345
4	0.7655	0.4992	18	0.7741	0.7874
5	0.9165	0.817	19	0.7752	0.7325
6	0.947	0.8985	20	0.7183	0.5715
7	0.8666	0.8779	21	0.7678	0.7693
8	0.8877	0.8862	22	0.8239	0.8006
9	0.5379	0.3555	23	0.8382	0.8454
10	0.8974	0.9172	24	0.7408	0.5912
11	0.5004	0.4653	25	0.9478	0.9085
12	0.7632	0.6728	26	0.9465	0.9013
13	0.8095	0.7551	27	0.3288	0.2955
14	0.7200	0.7426	28	0.2068	0.1526

Table 61: Correlations Between Ability Parameter Estimates Within and Across Models

	Dataset	
Parameters	A	B
$r(\widehat{\theta}_{1_{GMIRT}}, \widehat{\theta}_{1_{NC}})$	.93	.95
$r(\widehat{\theta}_{2_{GMIRT}}, \widehat{\theta}_{2_{NC}})$	.97	.88
$r(\widehat{\theta}_{1_{GMIRT}}, \widehat{\theta}_{2_{GMIRT}})$	.84	.70
$r(\widehat{\theta}_{1_{NC}}, \widehat{\theta}_{2_{NC}})$	.99	1.00
$r(ENC_{GMIRT}, ENC_{NC})$	.99	.99
$r(ENC_{GMIRT}, TOTAL_{GMIRT})$	.99	.99
$r(ENC_{NC}, TOTAL_{NC})$	.99	.99

Table 62: Mean Posterior Standard Deviations for Ability Estimates

Dataset	A		B	
Model	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
GMIRT	0.6356(.063)	0.9700(.104)	1.1260(.185)	0.7934(.128)
Noncompensatory	0.5448(.055)	0.4103(.038)	-.—-	-.—-
Compensatory	-.—-	-.—-	-.—-	-.—-

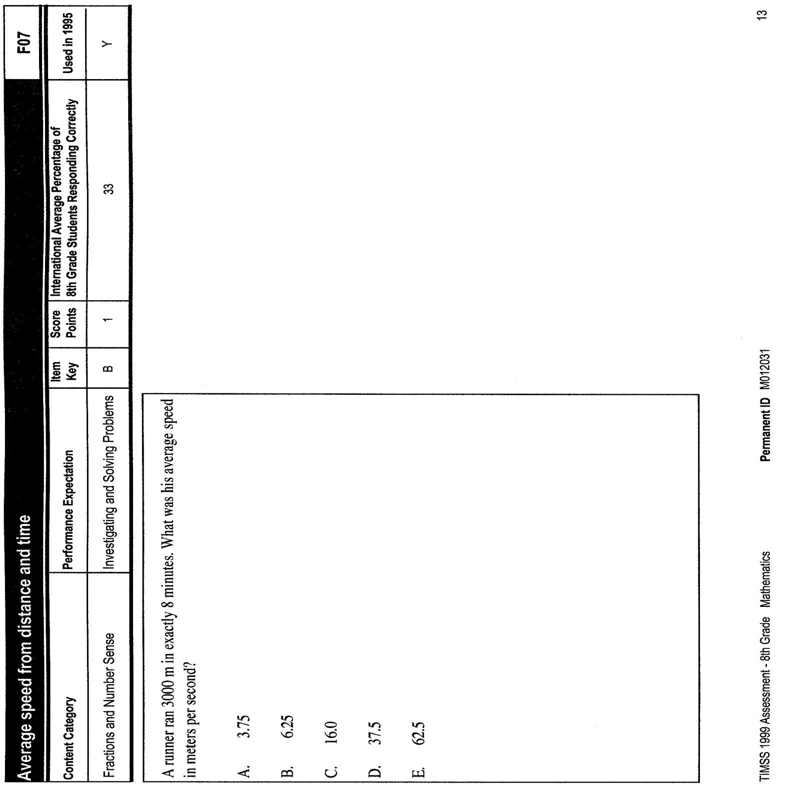
Table 63: Mean Proportion Reduction from Prior to Posterior Standard Deviations for Ability Estimates

Dataset	A		B	
Model	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
GMIRT	0.3643(.063)	0.0300(.104)	-.1260(.185)	0.2066(.128)
Noncompensatory	0.4553(.055)	0.5897(.038)	-.—-	-.—-
Compensatory	-.—-	-.—-	-.—-	-.—-

Table 64: Posterior Means for Compensation Parameters and Working Memory Scores for BAS Matrices

Item	$\hat{\mu_c}$,Dataset A	$\hat{\mu_c}$,Dataset B	Number of Rules	Weighted Score
1	-.—-	-.—-	1	1
2	0.1109	0.0676	1	1
3	0.0642	0.1088	1	2
4	0.0879	0.1661	1	2
5	0.0326	0.0593	1	4
6	0.0172	0.0518	2	5
7	0.1029	0.0868	2	4
8	0.0607	0.059	1	4
9	0.3677	0.5178	2	5
10	0.0563	0.0403	1	2
11	0.7476	0.7162	2	2
12	0.2286	0.2901	2	8
13	0.1612	0.2046	1	2
14	0.2532	0.2147	3	3
15	0.1417	0.1846	1	2
16	0.1891	0.1525	2	5
17	0.3301	0.3036	2	6
18	0.1892	0.1456	3	12
19	0.1971	0.2263	2	8
20	0.2487	0.3705	2	8
21	0.1838	0.1499	1	3
22	0.1231	0.1389	3	6
23	0.1125	0.0902	1	3
24	0.1479	0.2565	3	6
25	0.0106	0.0137	4	20
26	0.0107	0.0124	4	20
27	0.6069	0.3628	3	9
28	0.5103	0.4852	4	11

Appendix B. Figures



Average speed from distance and time

F07

Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Fractions and Number Sense	Investigating and Solving Problems	B	1	33	Y

A runner ran 3000 m in exactly 8 minutes. What was his average speed in meters per second?

A. 3.75

B. 6.25

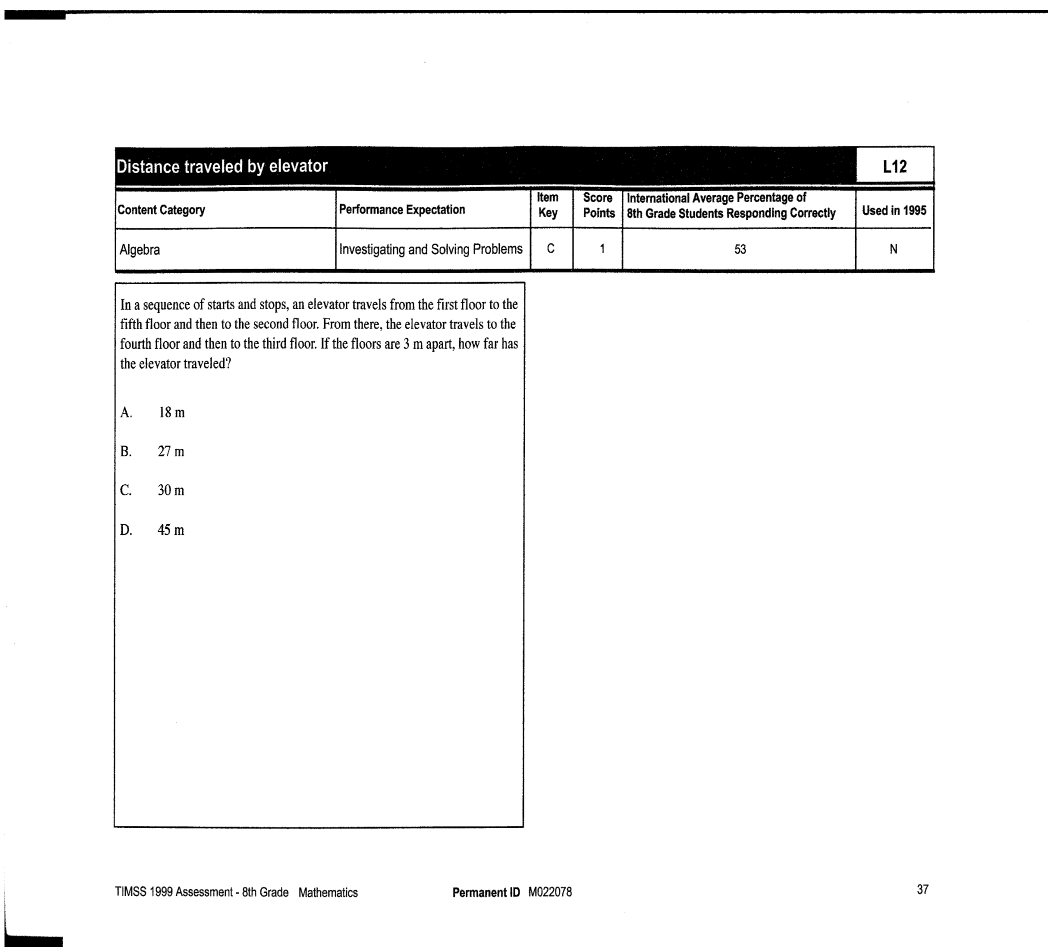
C. 16.0

D. 37.5

E. 62.5

TIMSS 1999 Assessment - 8th Grade Mathematics Permanent ID M012031 13

Figure 1: Mathematics problem with hypothesized compensatory Structure



Distance traveled by elevator					L12
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Algebra	Investigating and Solving Problems	C	1	53	N

In a sequence of starts and stops, an elevator travels from the first floor to the fifth floor and then to the second floor. From there, the elevator travels to the fourth floor and then to the third floor. If the floors are 3 m apart, how far has the elevator traveled?

A. 18 m

B. 27 m

C. 30 m

D. 45 m

TIMSS 1999 Assessment - 8th Grade Mathematics **Permanent ID** M022078 37

Figure 2: Mathematics problem with hypothesized noncompensatory Structure

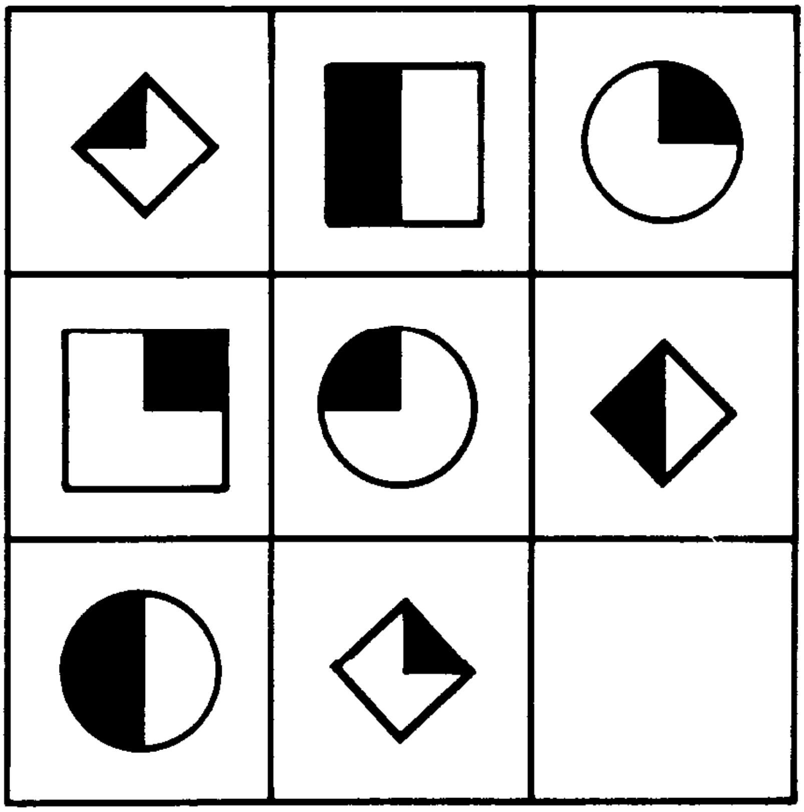


Figure 3: A matrix completion task

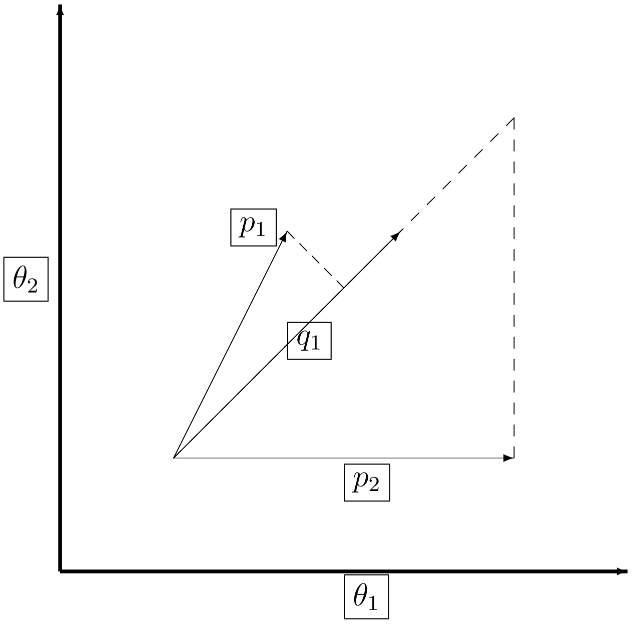


Figure 4: Stimulus Compensatory model adapted from (Coombs, 1964)

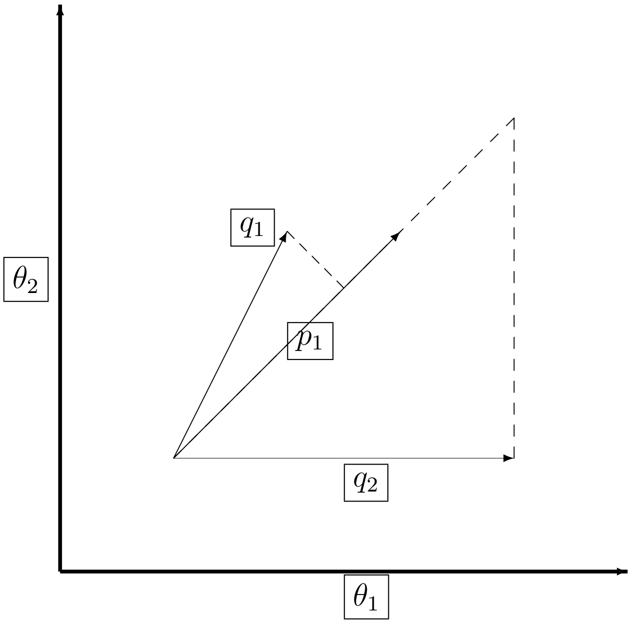


Figure 5: Individual compensatory model (adapted from Coombs, 1964)

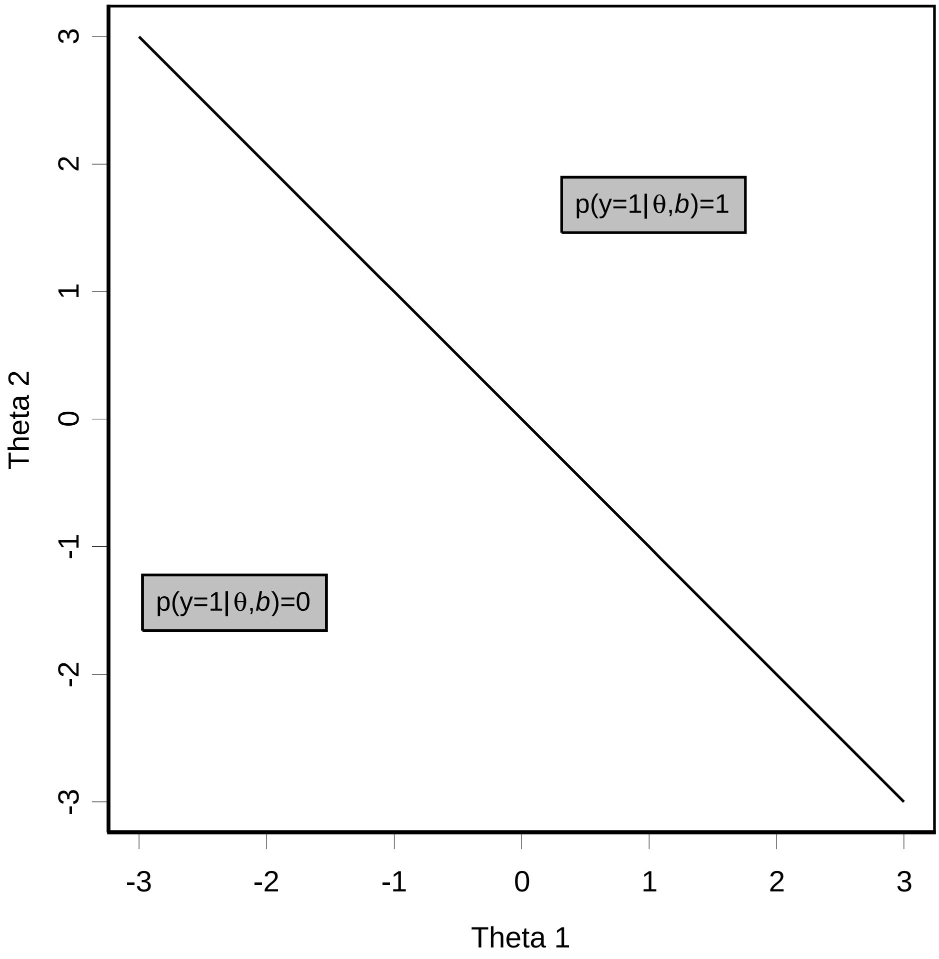


Figure 6: Threshold for deterministic compensatory Model

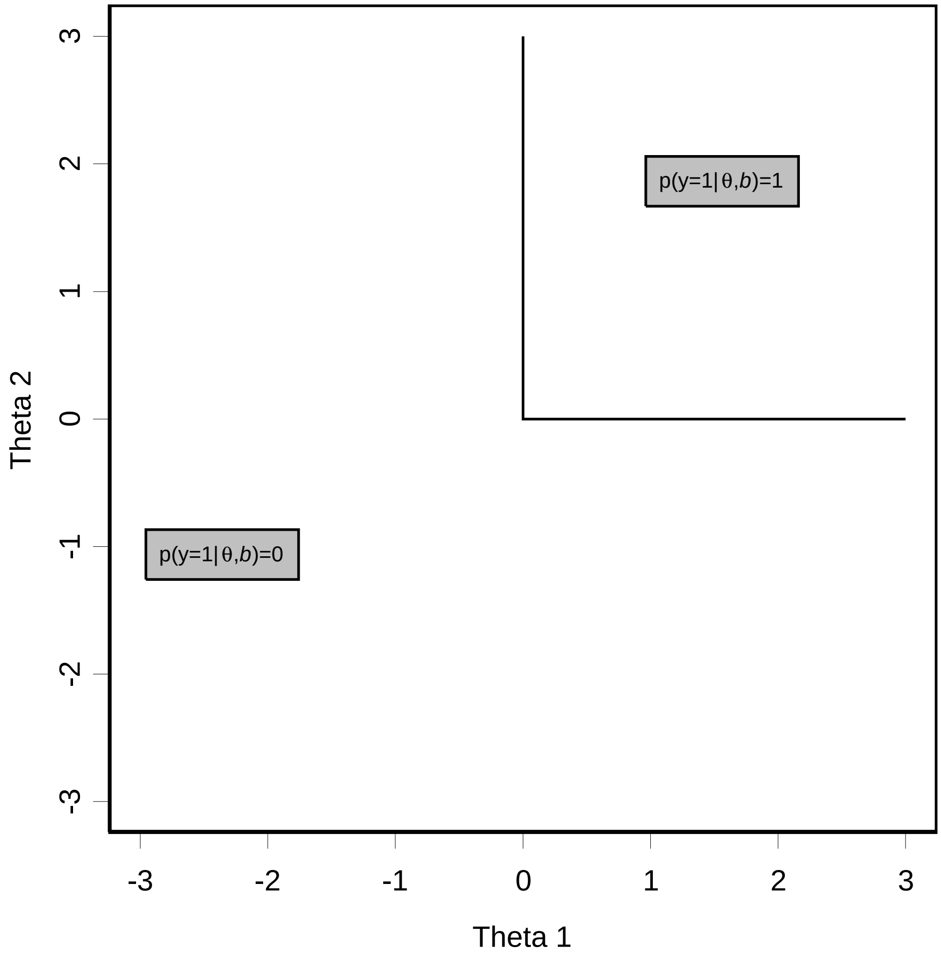


Figure 7: Threshold for deterministic noncompensatory Model

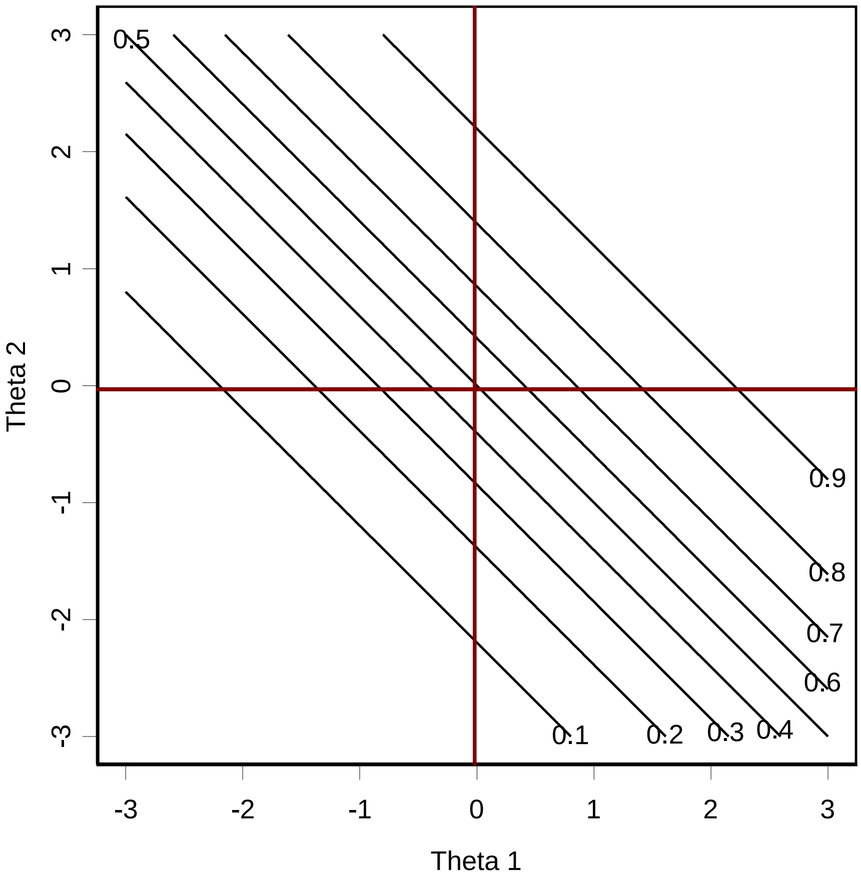


Figure 8: Probability contour plot for two-dimensional compensatory Rasch model

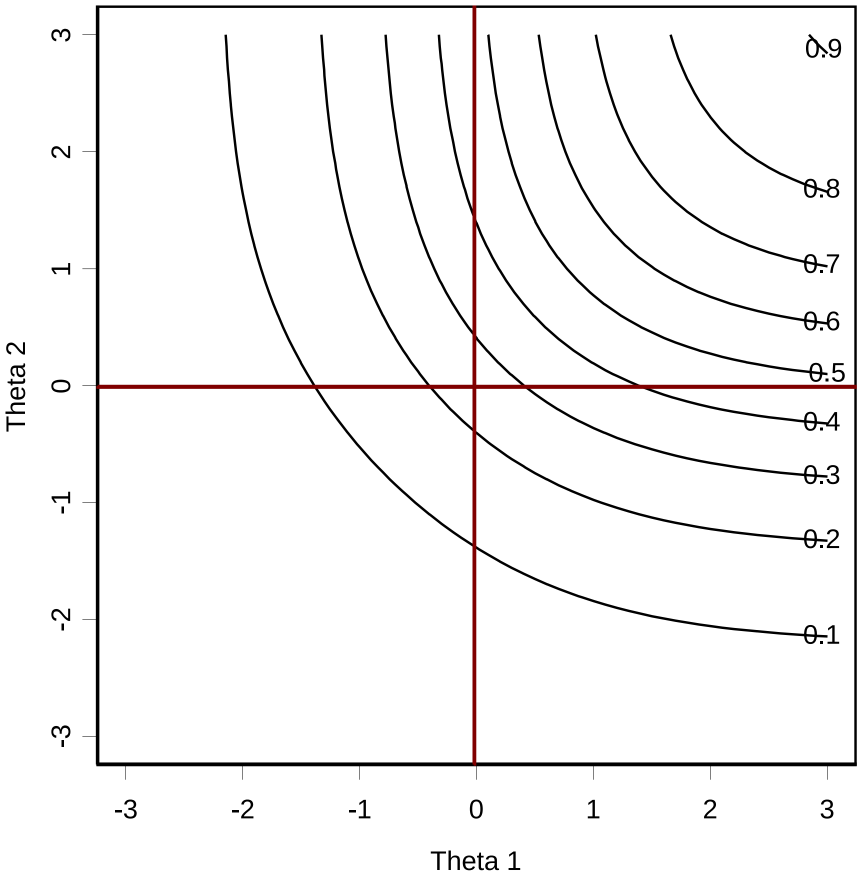


Figure 9: Probability contour plot for two-dimensional noncompensatory Rasch model

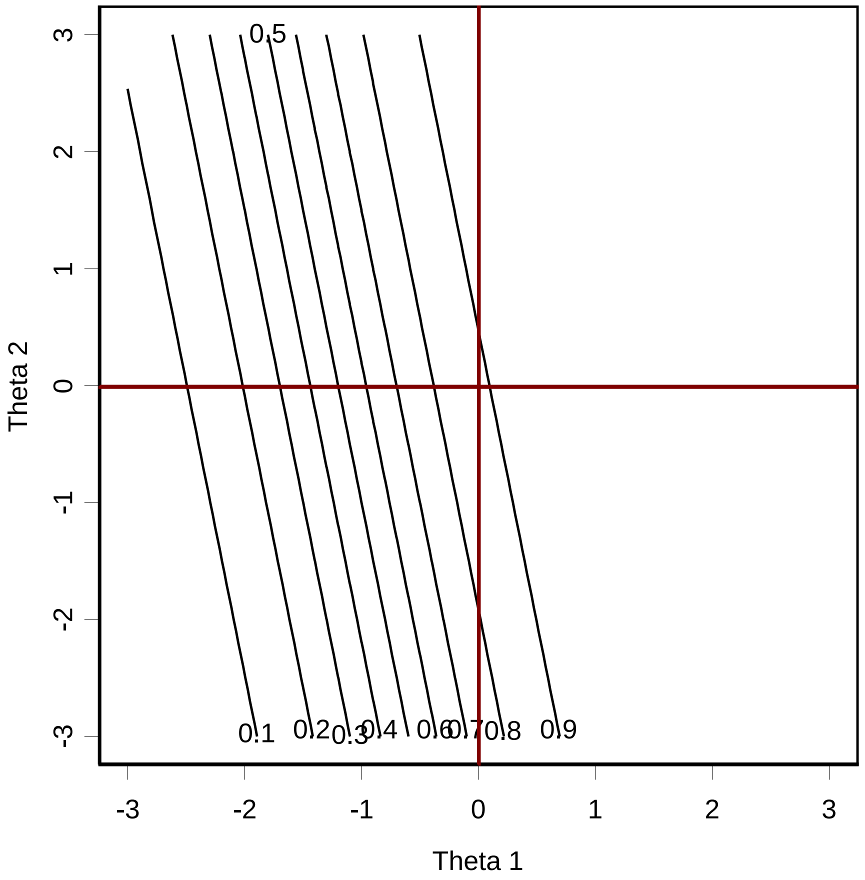


Figure 10: Probability contour plot for compensatory item, dimension 1 predominating

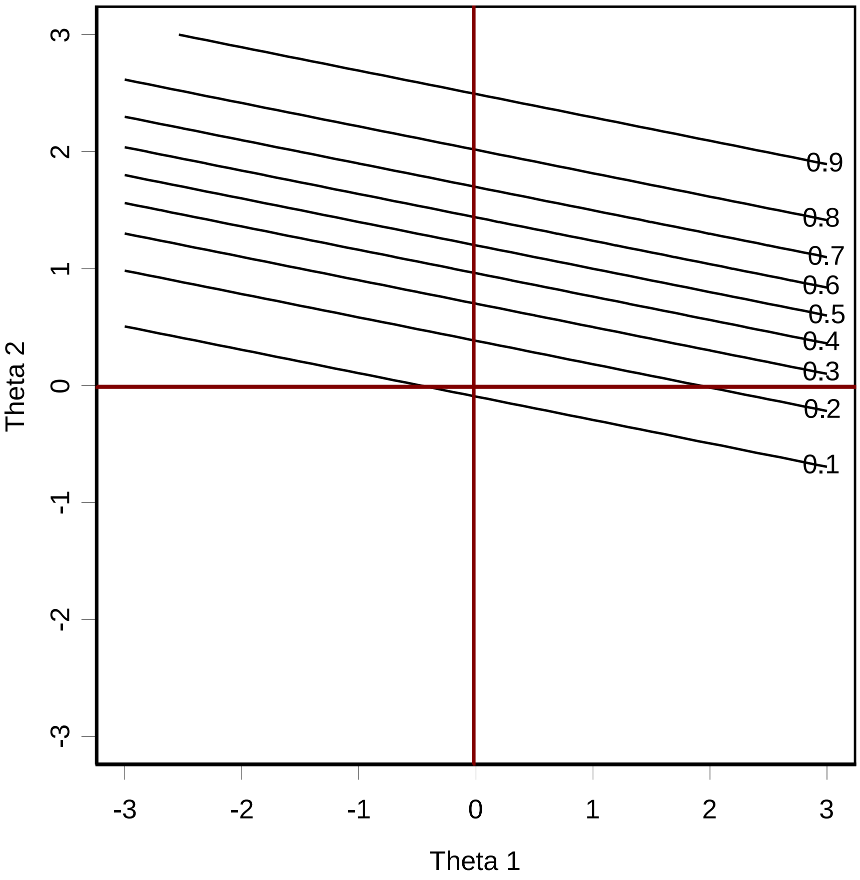


Figure 11: Probability contour plot for compensatory item, dimension 2 predominating

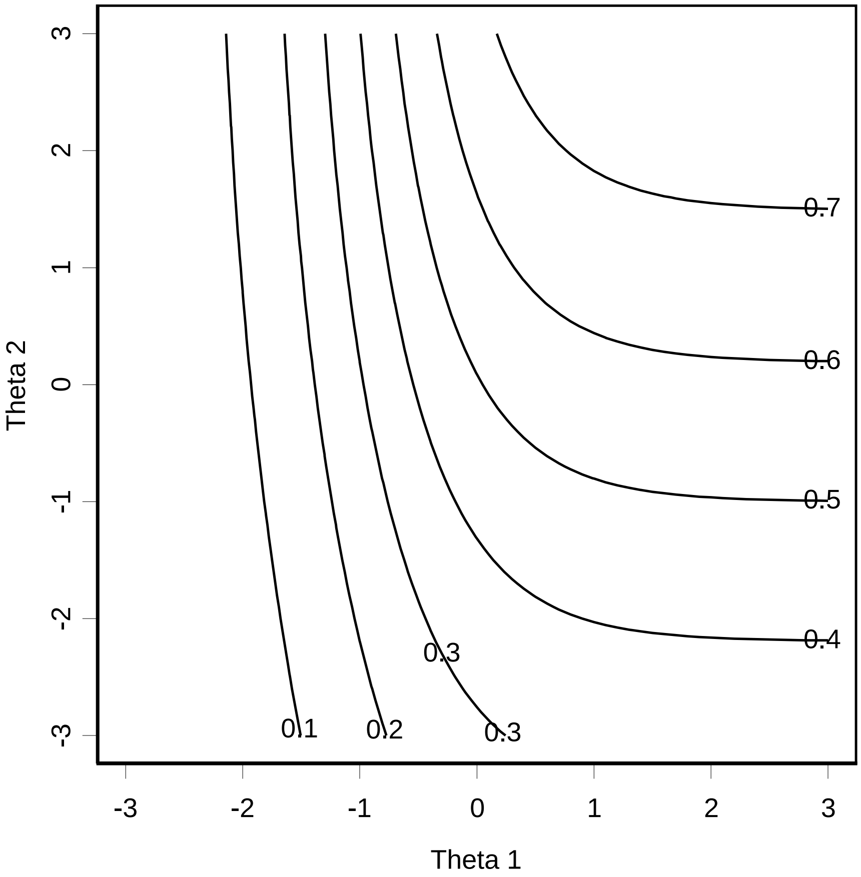


Figure 12: Probability contour plot for noncompensatory item, dimension 1 predominating

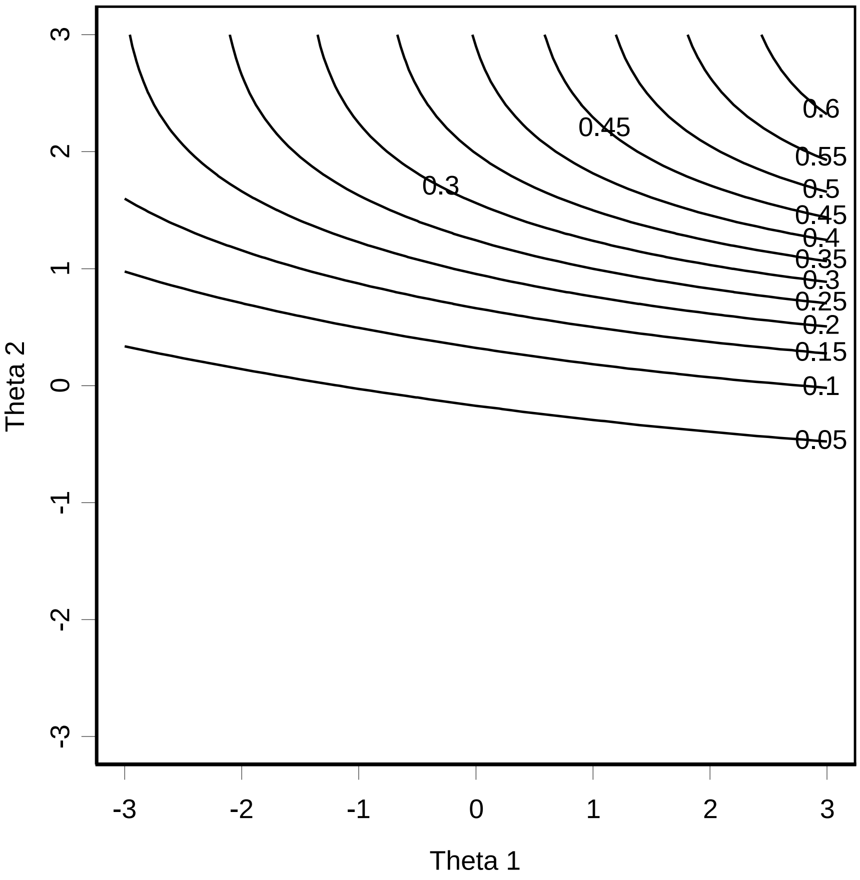


Figure 13: Probability contour plot for noncompensatory item, dimension 2 predominating

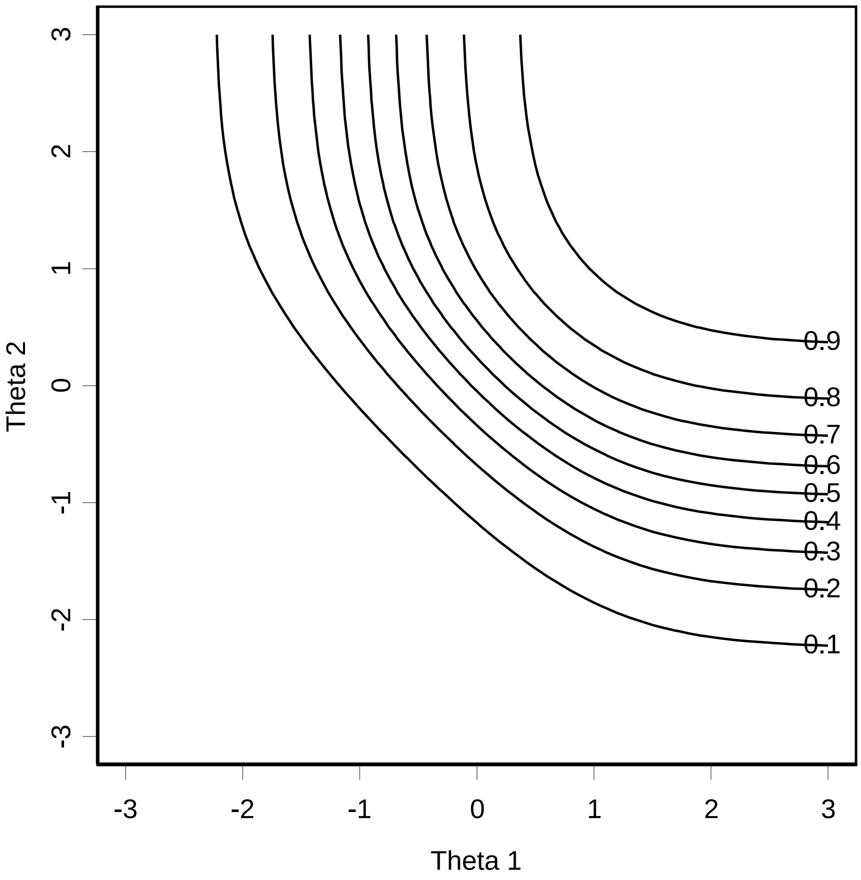


Figure 14: Probability contour plot for GMIRT item, μ_c = .20

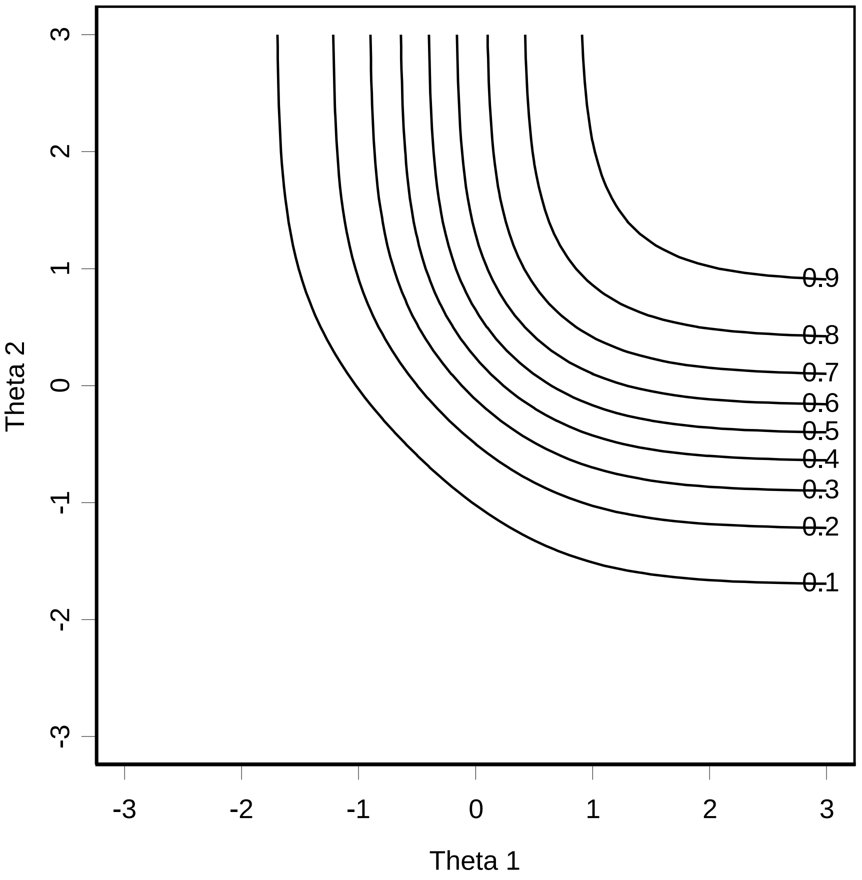


Figure 15: Probability contour plot for GMIRT item, μ_c = .50

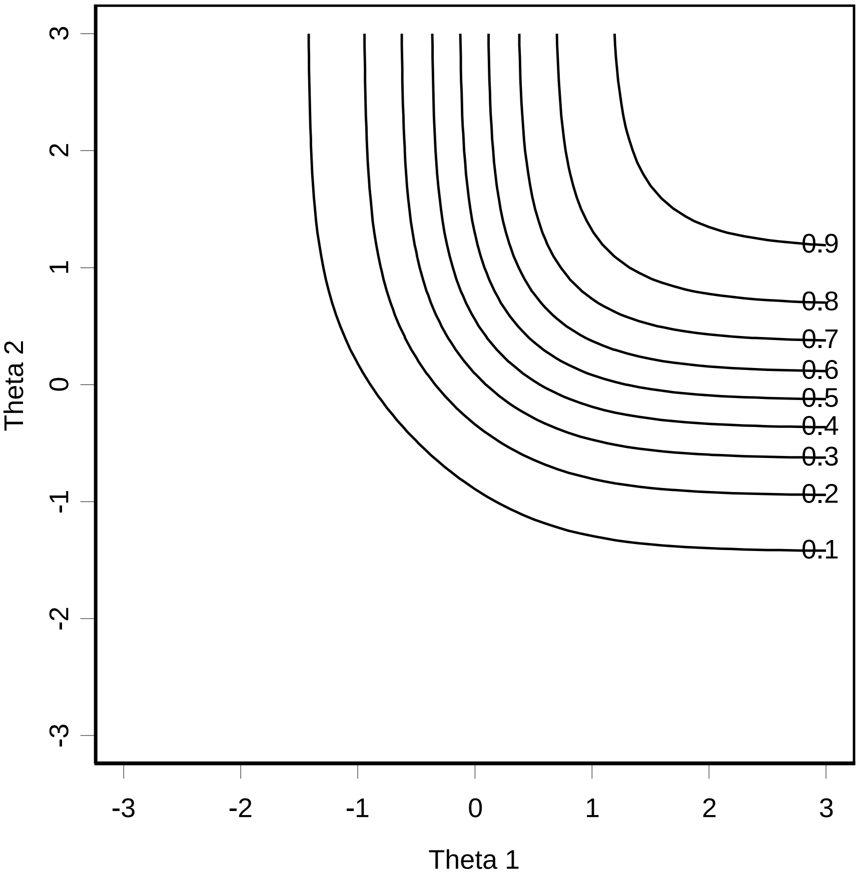


Figure 16: Probability contour plot for GMIRT item, μ_c = .80

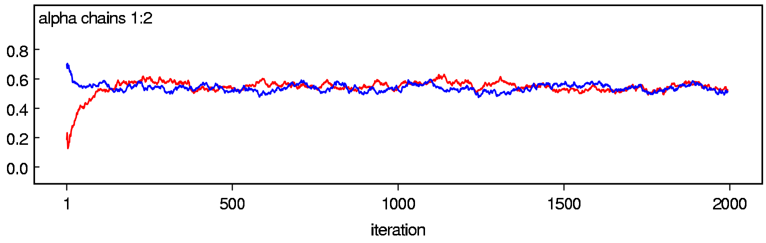


Figure 17: MCMC history plot: ”Beeline” for convergence

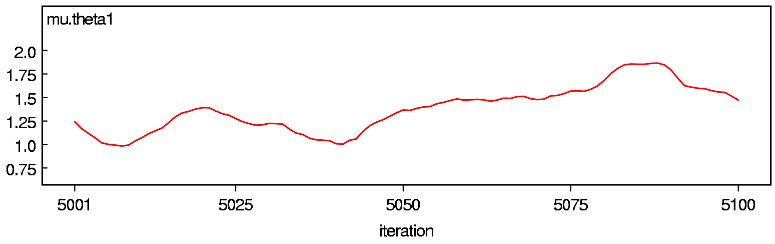


Figure 18: MCMC history plot: ”Wandering”

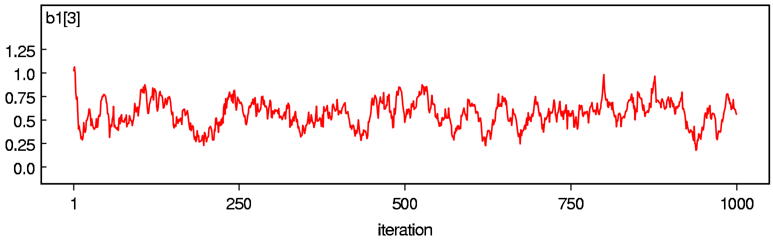


Figure 19: MCMC history plot: ”Snaking”

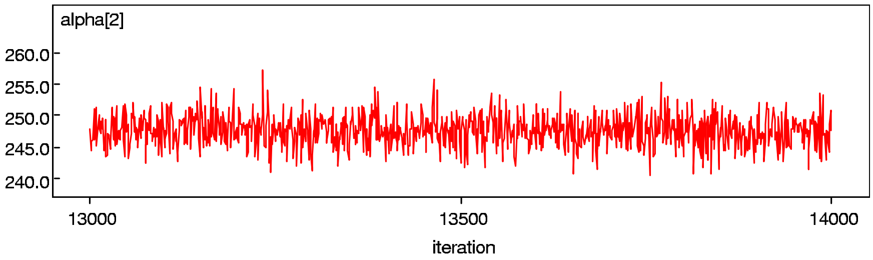
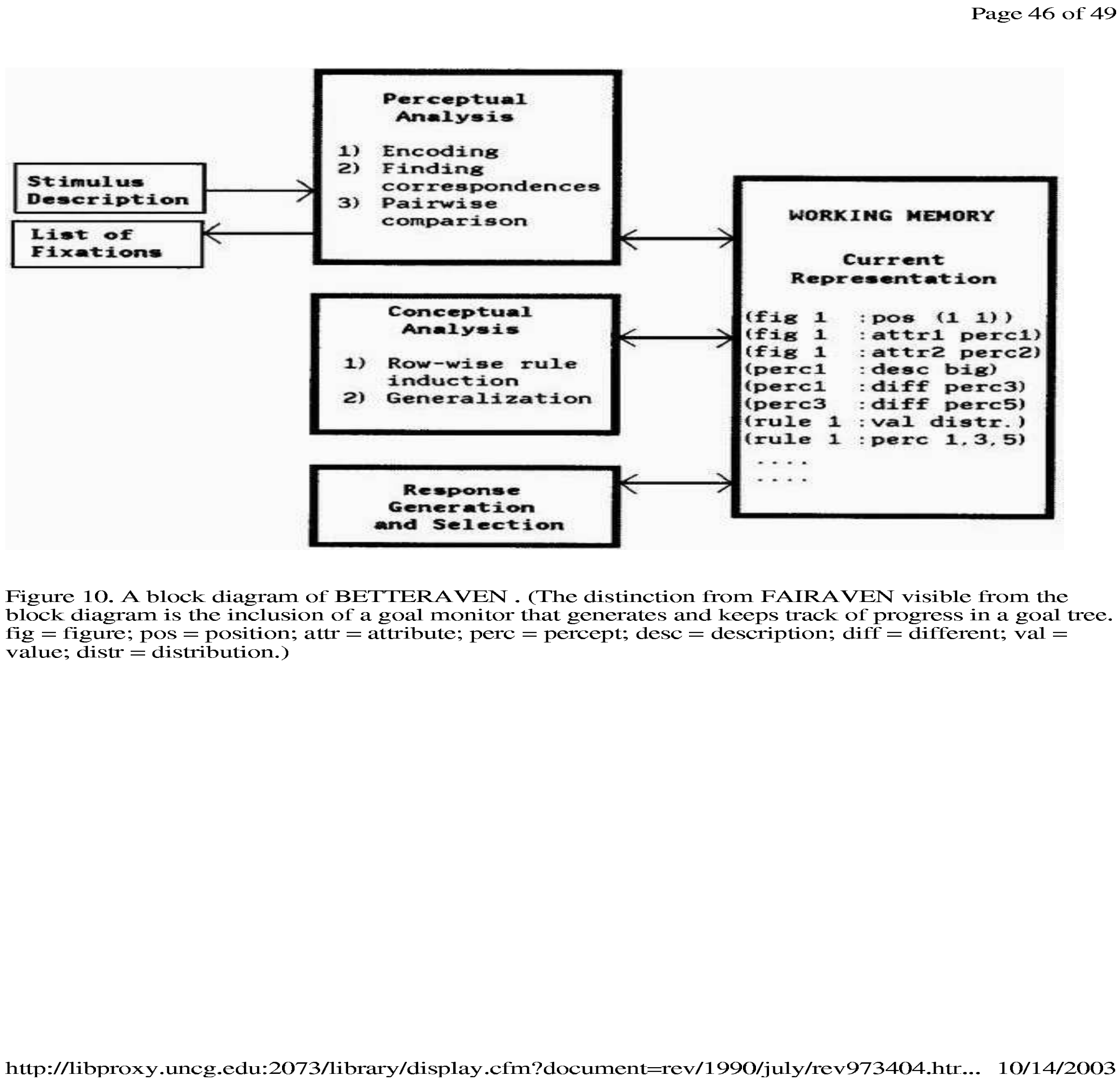


Figure 20: MCMC history plot: Ideal sampling pattern



Page 46 of 49

Figure 10. A block diagram of BETTERAVEN . (The distinction from FAIRAVEN visible from the block diagram is the inclusion of a goal monitor that generates and keeps track of progress in a goal tree. fig = figure; pos = position; attr = attribute; perc = percept; desc = description; diff = different; val = value; distr = distribution.)

http://libproxy.uncg.edu:2073/library/display.cfm?document=rev/1990/july/rev973404.htr... 10/14/2003

Figure 21: FAIRAVEN model for performance on Ravens Matrices (Carpenter, Just, & Shell, 1990)

Page 47 of 49

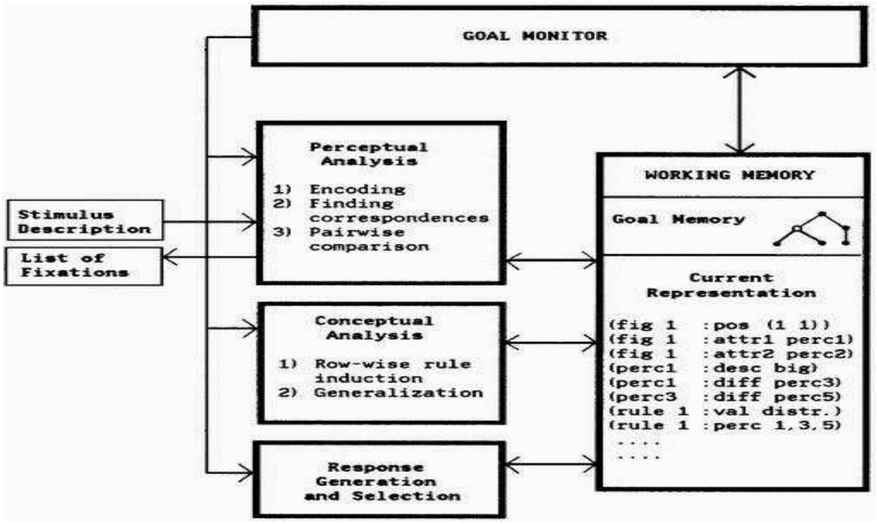


Figure 11. The elapsed time from the beginning of the trial to the verbal description of each of the rules in a problem.

http://libproxy.uncg.edu:2073/library/display.cfm?document=rev/1990/july/rev973404.htr... 10/14/2003

Figure 22: BETTERAVEN model for performance on Ravens Matrices (Carpenter, Just, & Shell, 1990)

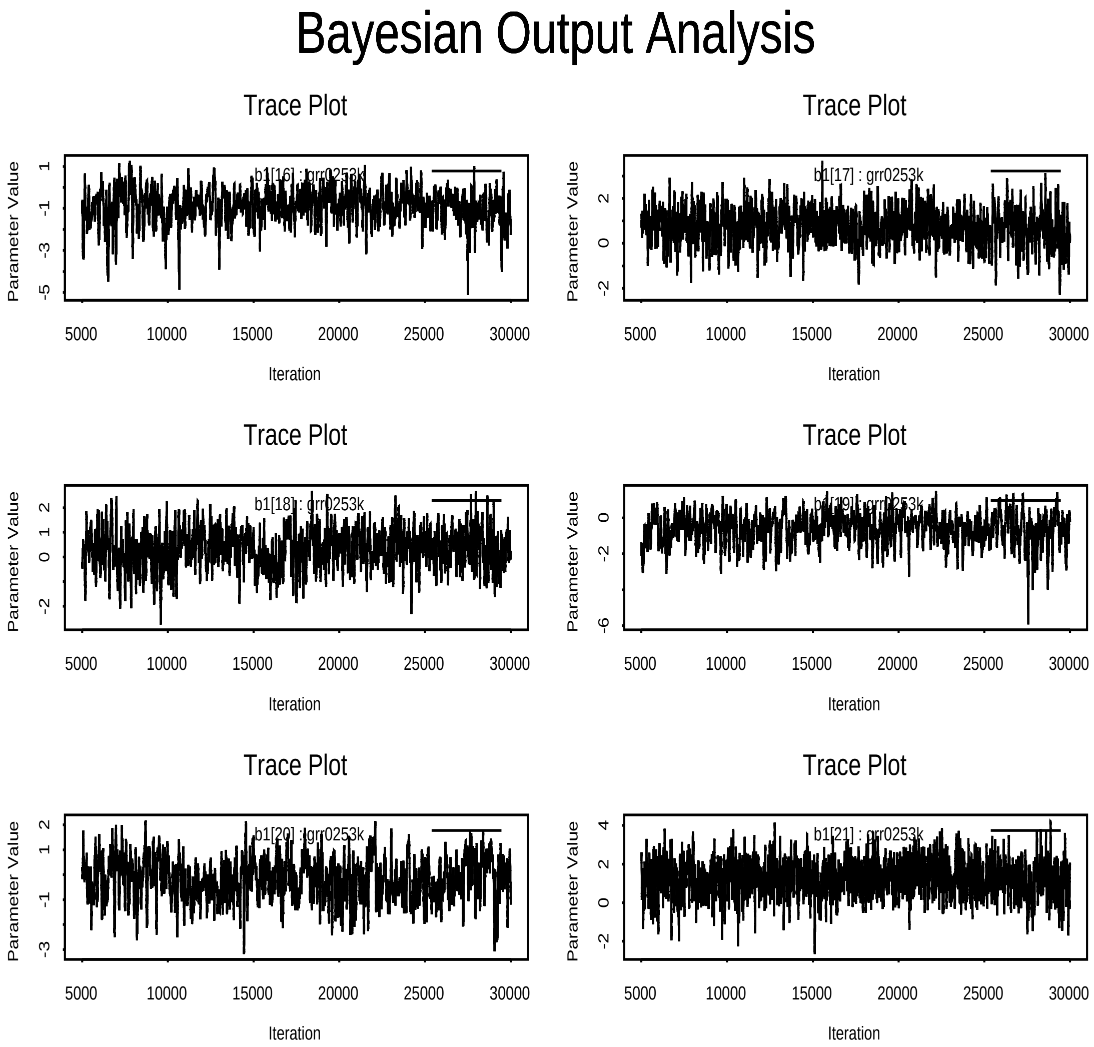


Figure 23: Free GMIRT estimation: Sampling histories for selected b parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

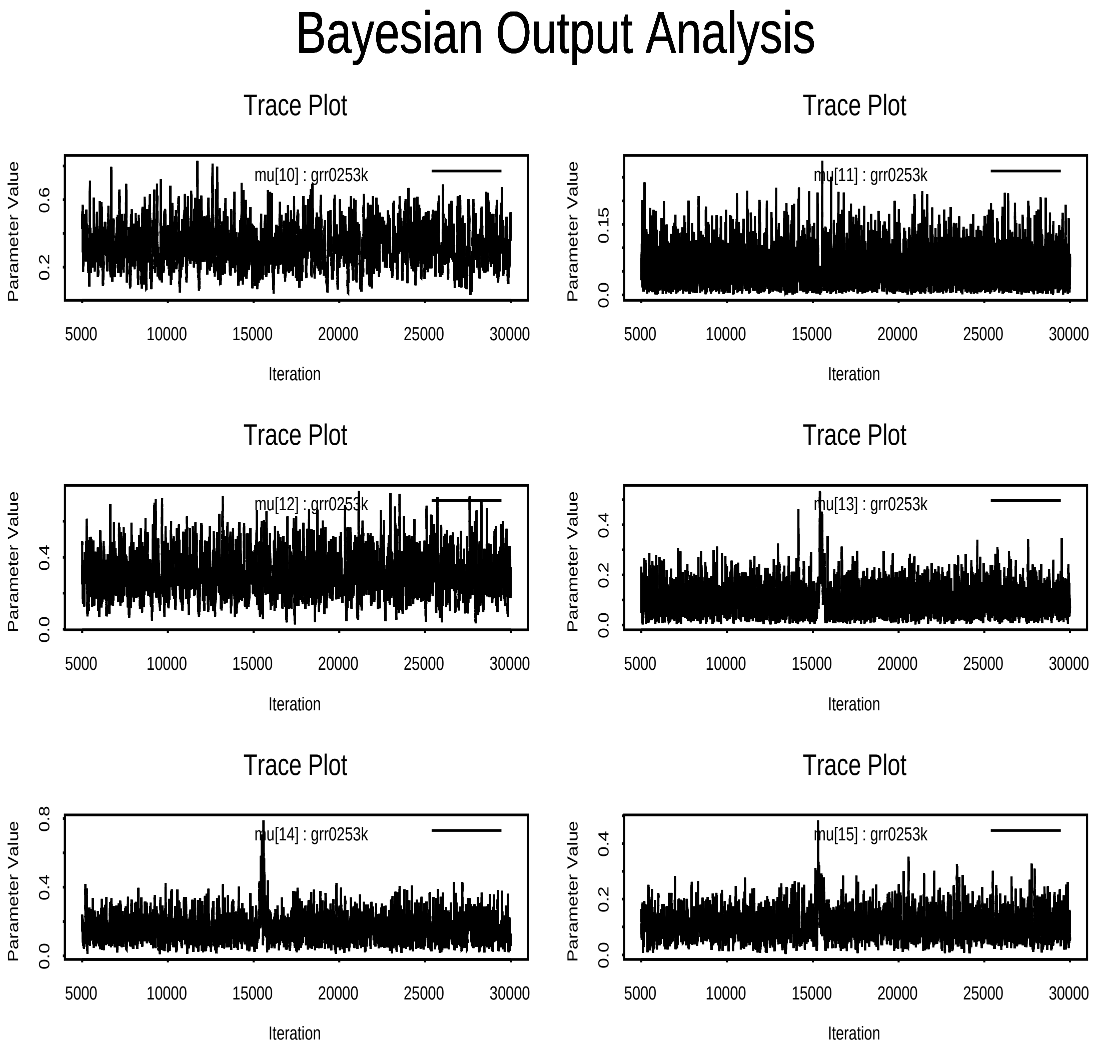


Figure 24: Free GMIRT estimation: Sampling histories for selected μ_c parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

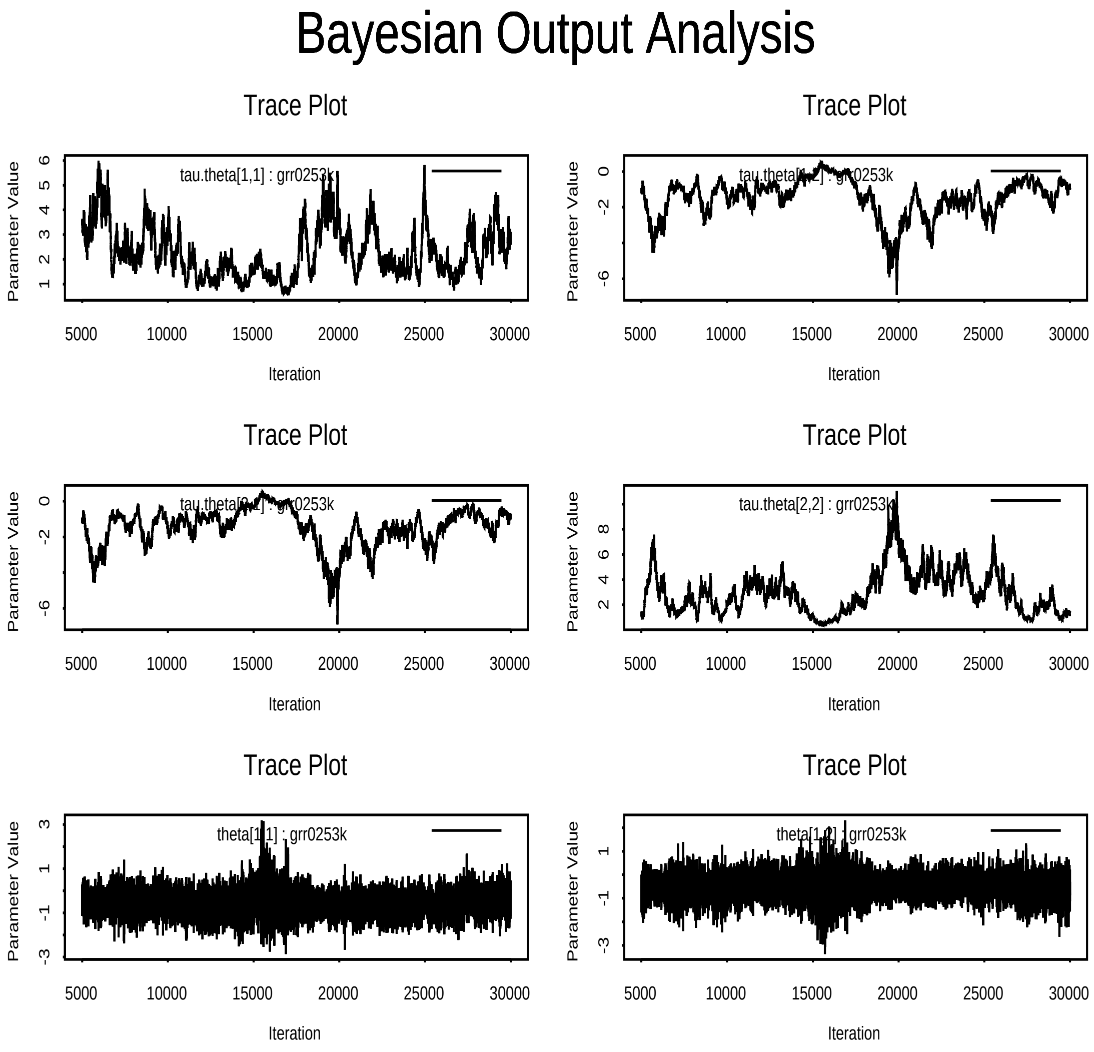


Figure 25: Free GMIRT estimation: Sampling histories for τ_θ and selected θ parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

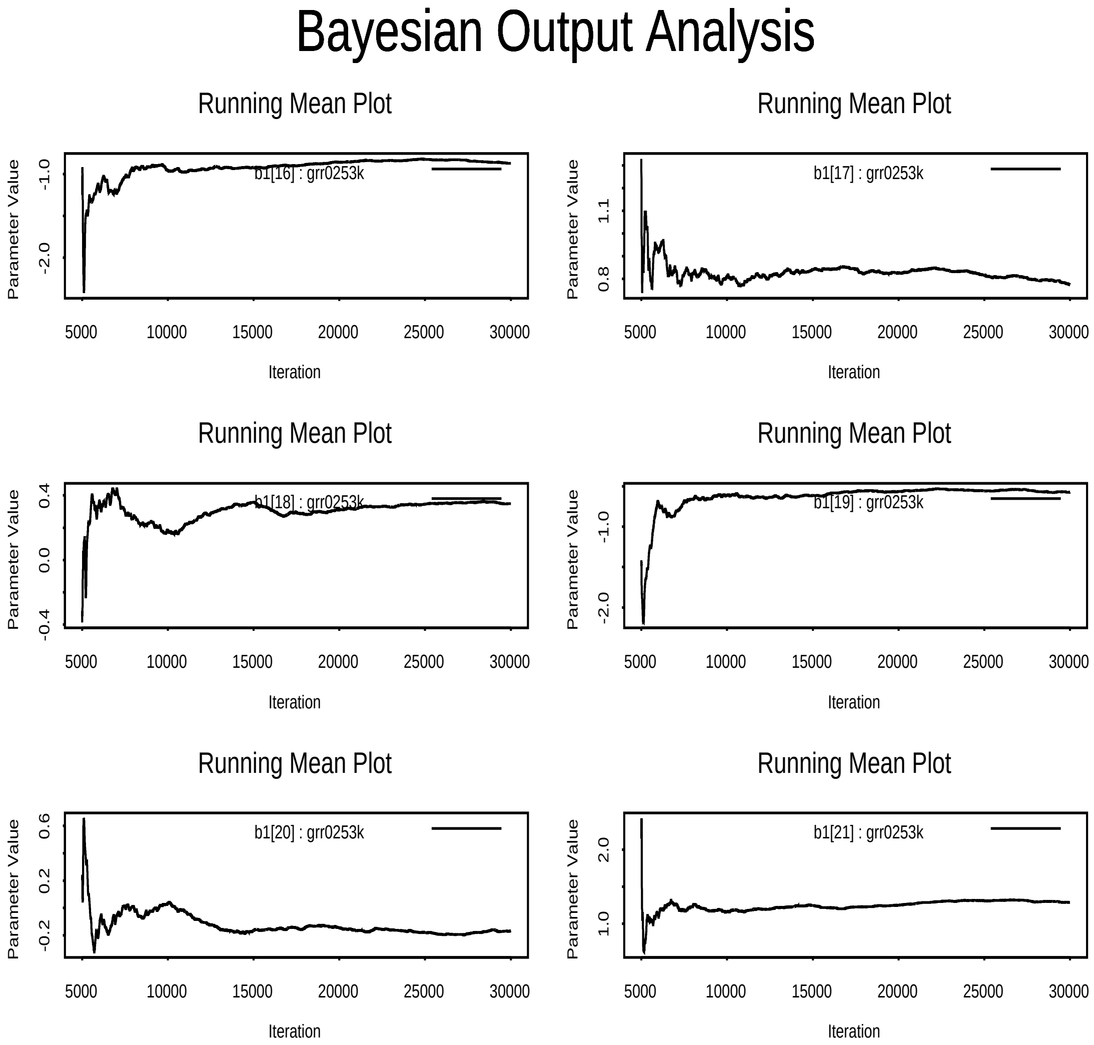


Figure 26: Free GMIRT estimation: Selected running mean plots for b parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

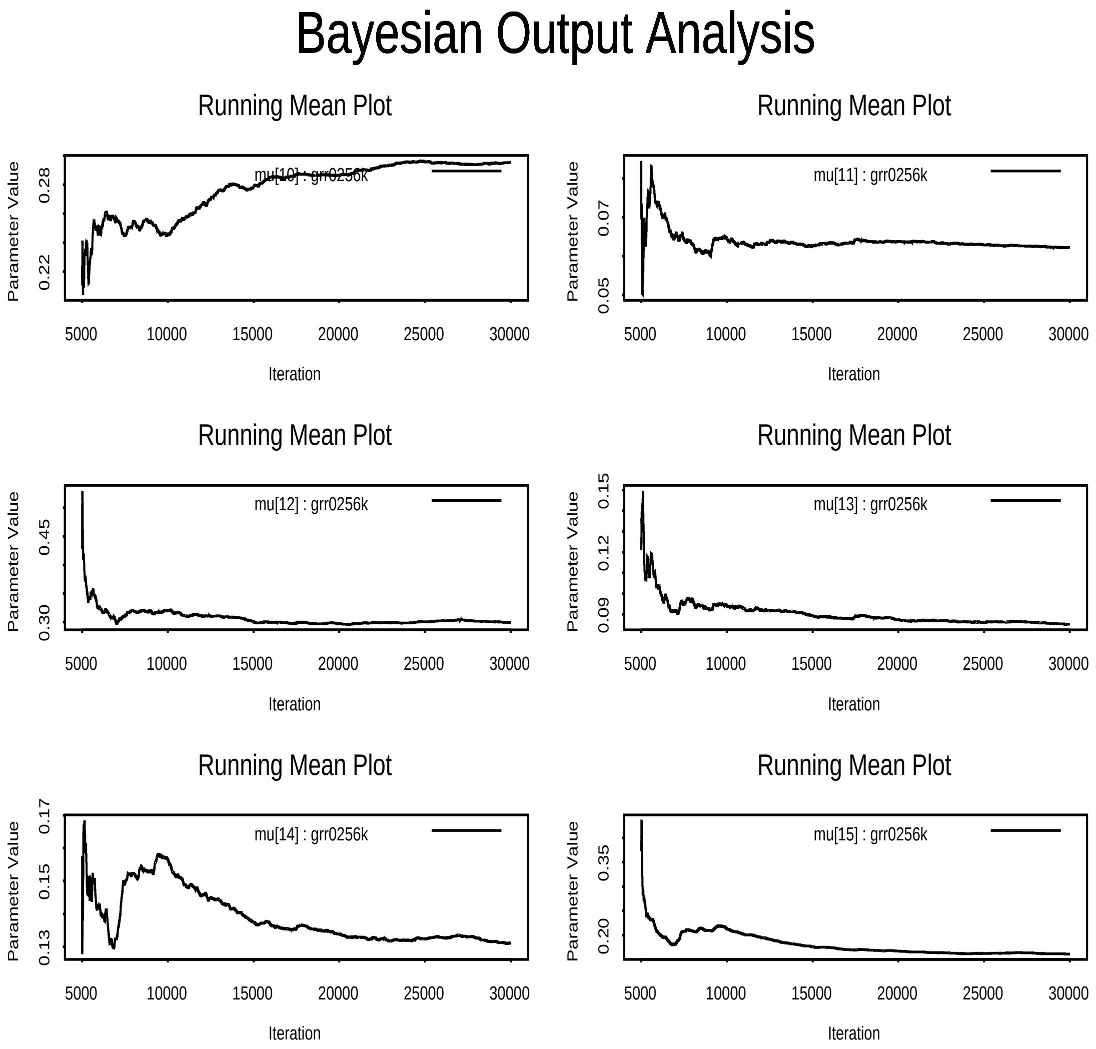


Figure 27: Free GMIRT estimation: Selected running mean plots for μ_c parameters, N = 6000, $r(\theta_1, \theta_2) = 0$

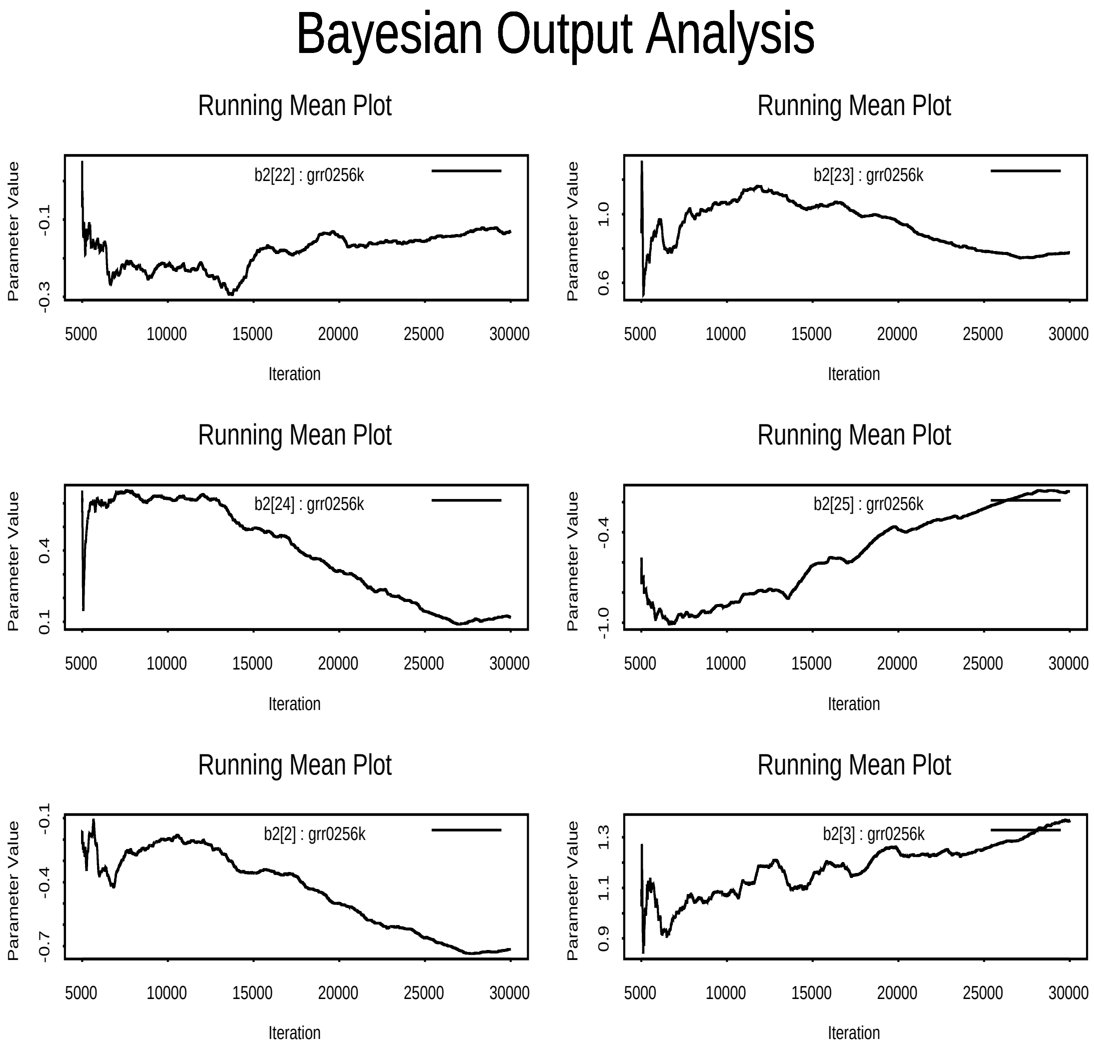


Figure 28: Free GMIRT estimation: Selected running mean plots for b_1 parameters, N = 6000, $r(\theta_1, \theta_2) = 0$

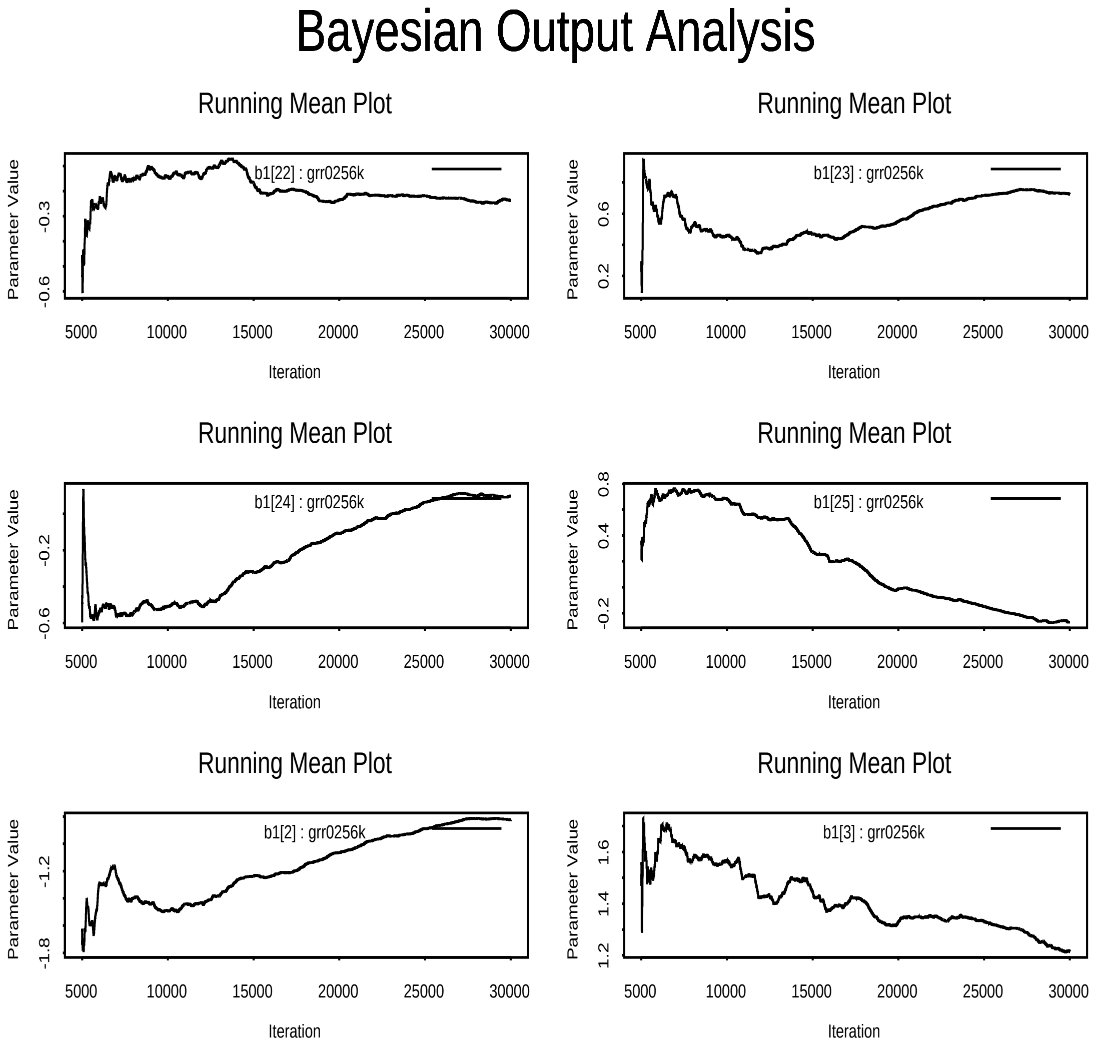


Figure 29: Free GMIRT estimation: Selected running mean plots for b_2 parameters, N = 6000, $r(\theta_1, \theta_2) = 0$

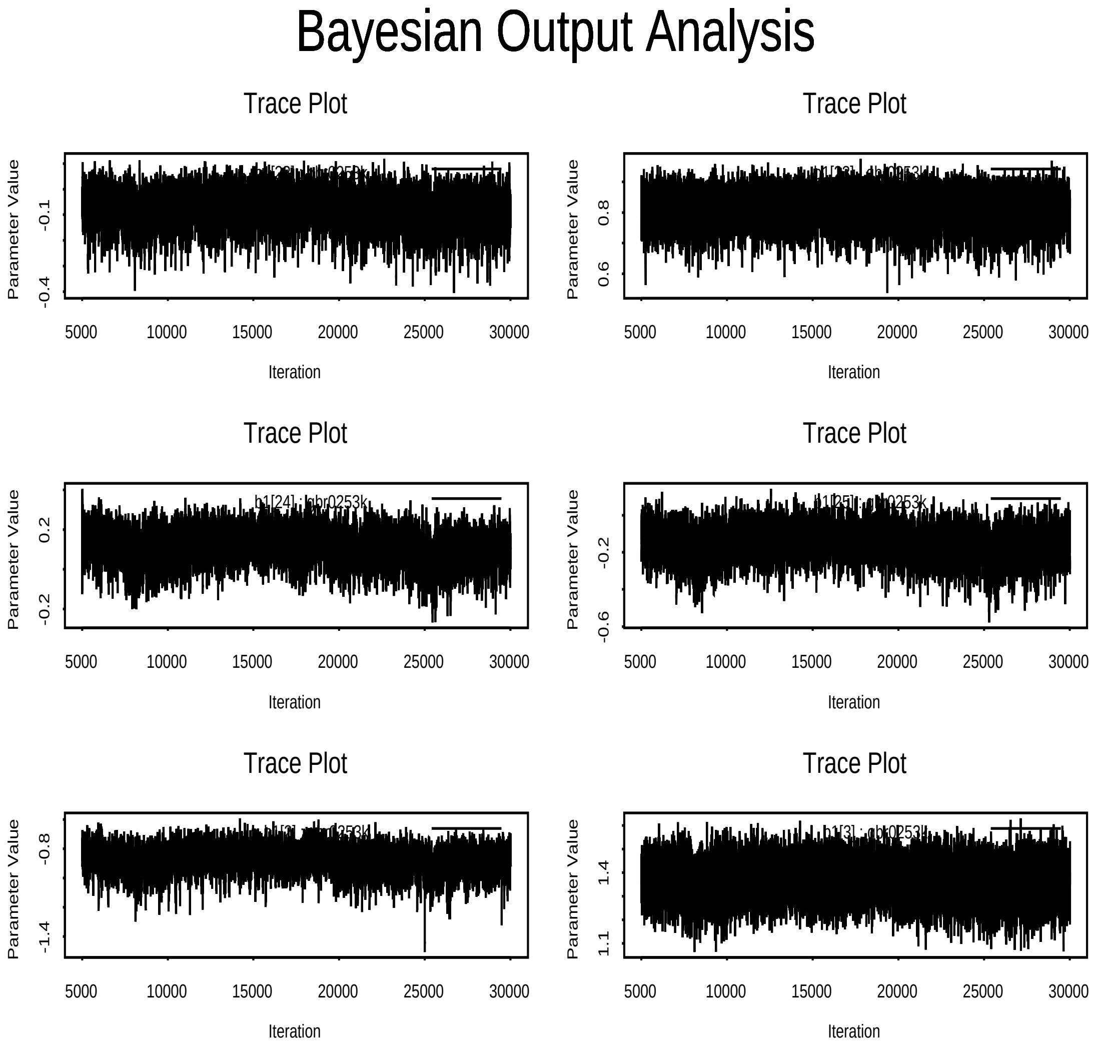


Figure 30: Constrained GMIRT estimation: Sampling histories for selected $\bar{b}$ parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

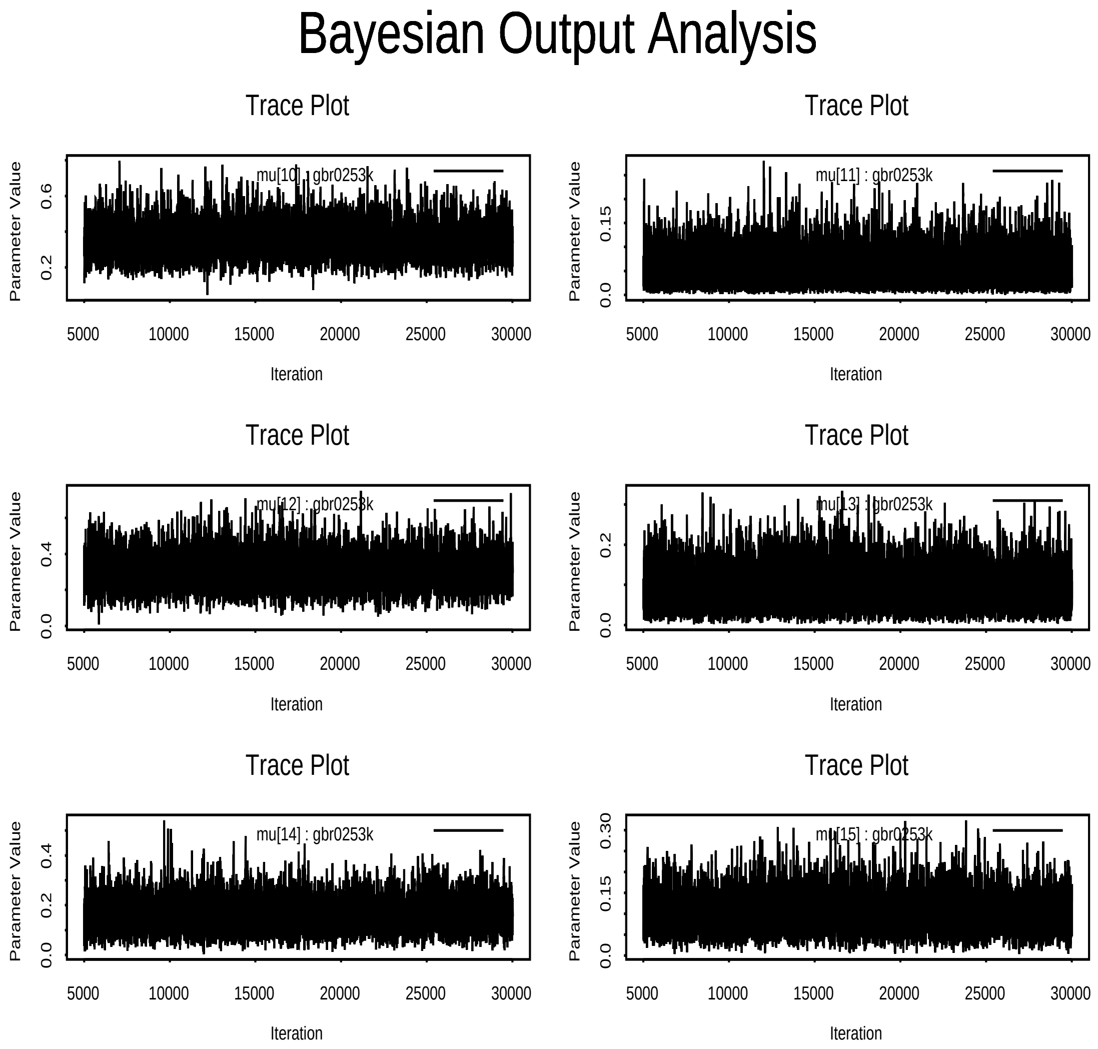


Figure 31: Constrained GMIRT estimation: Sampling histories for selected μ_c parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

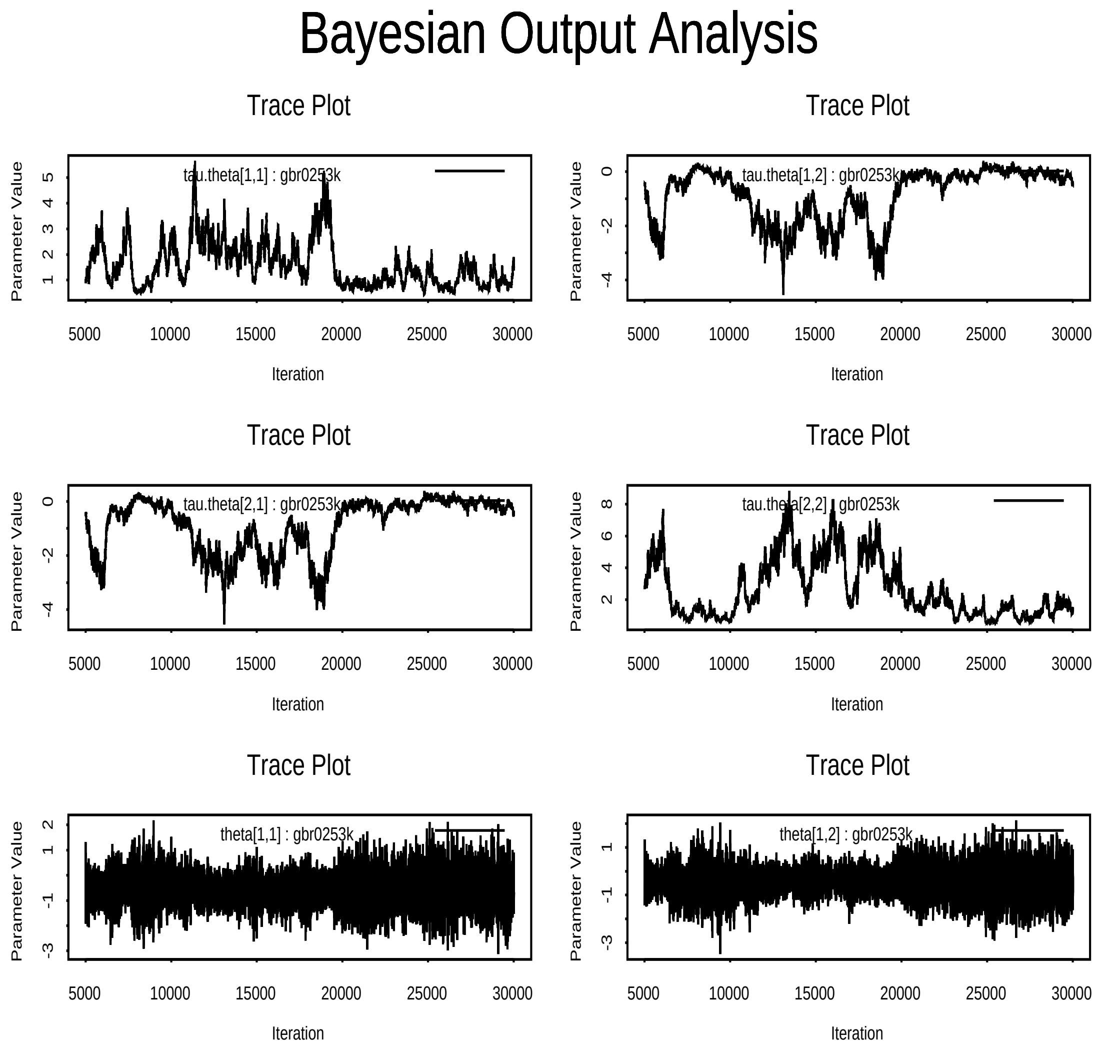


Figure 32: Constrained GMIRT estimation: Sampling histories for τ_θ and selected θ parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

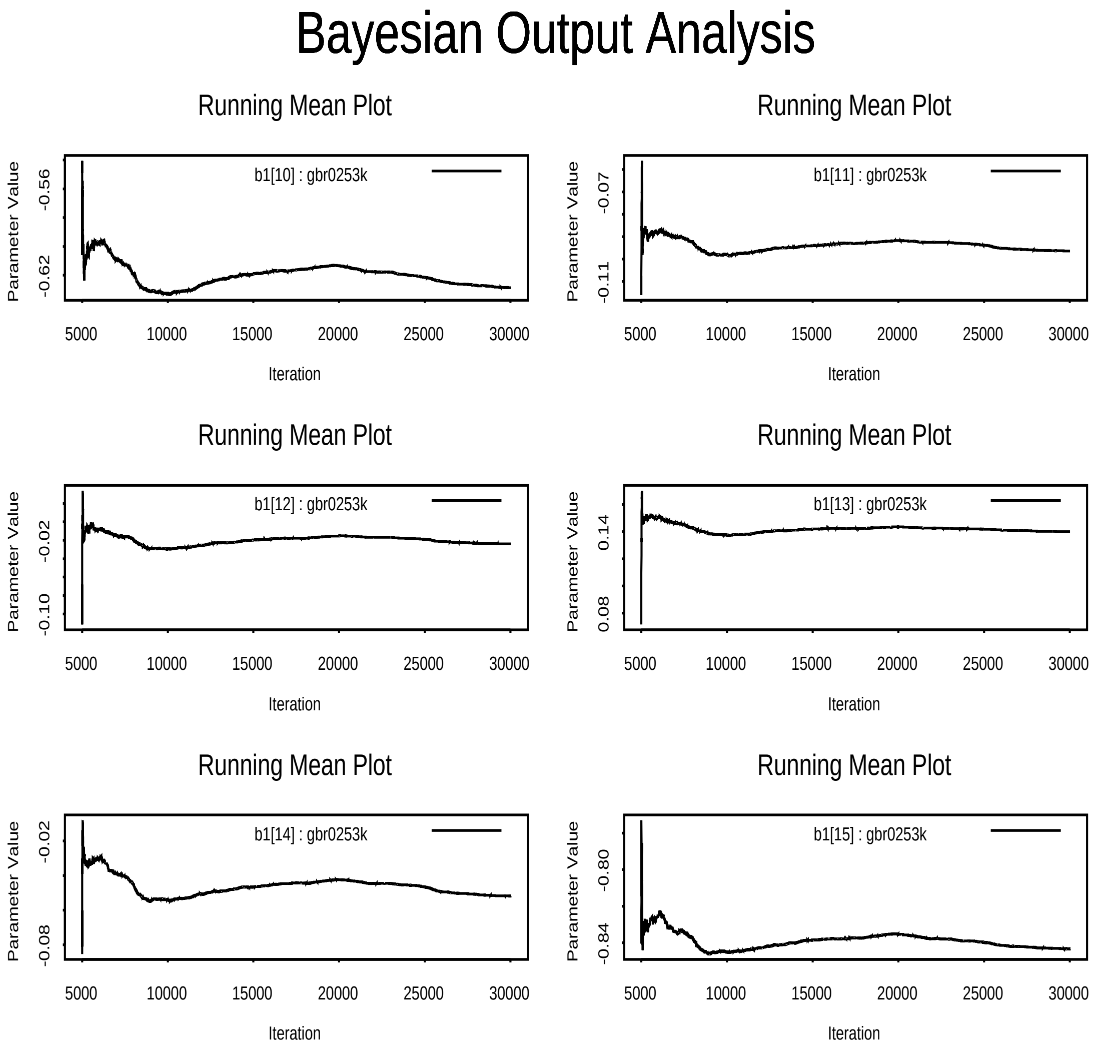


Figure 33: Constrained GMIRT estimation: Selected running mean plots for $\bar{b}$ parameters, N = 3000, $r(\theta_1, \theta_2) = 0$

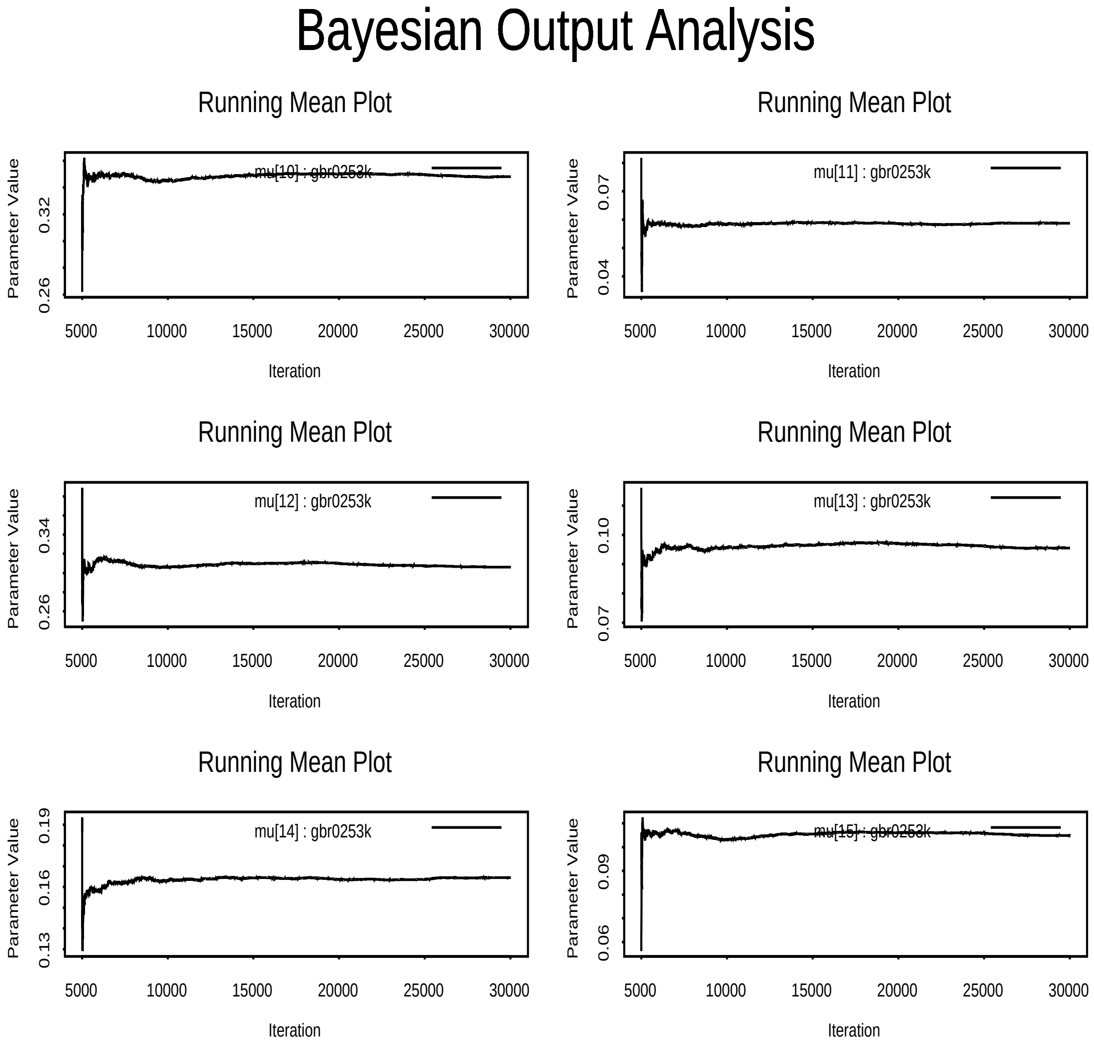


Figure 34: Constrained GMIRT estimation: Selected running mean plots for μ_c parameters, N = 6000, $r(\theta_1, \theta_2) = 0$

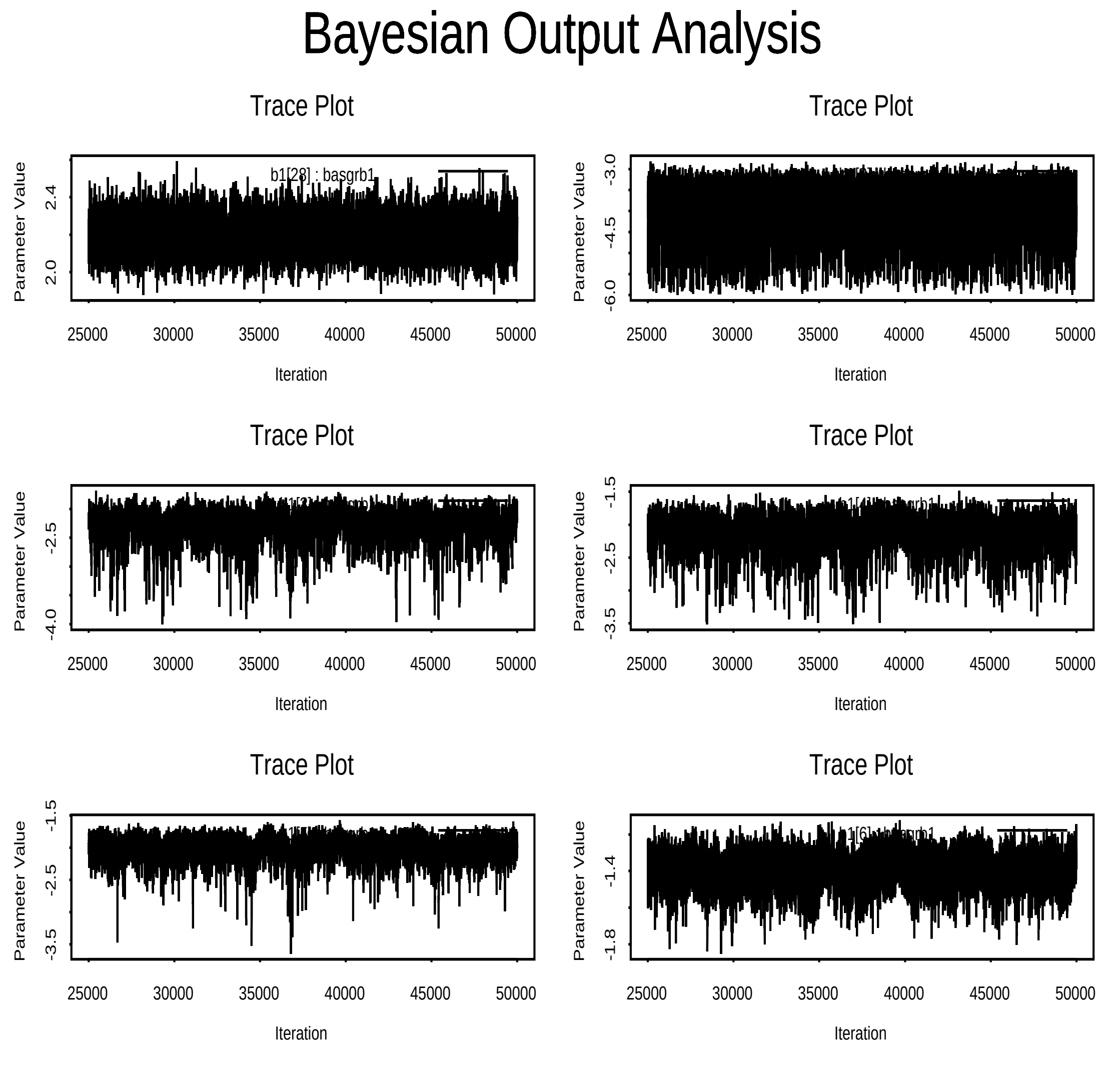


Figure 35: Application study, sampling histories: Selected $\bar{b}$ parameters, GMIRT Model

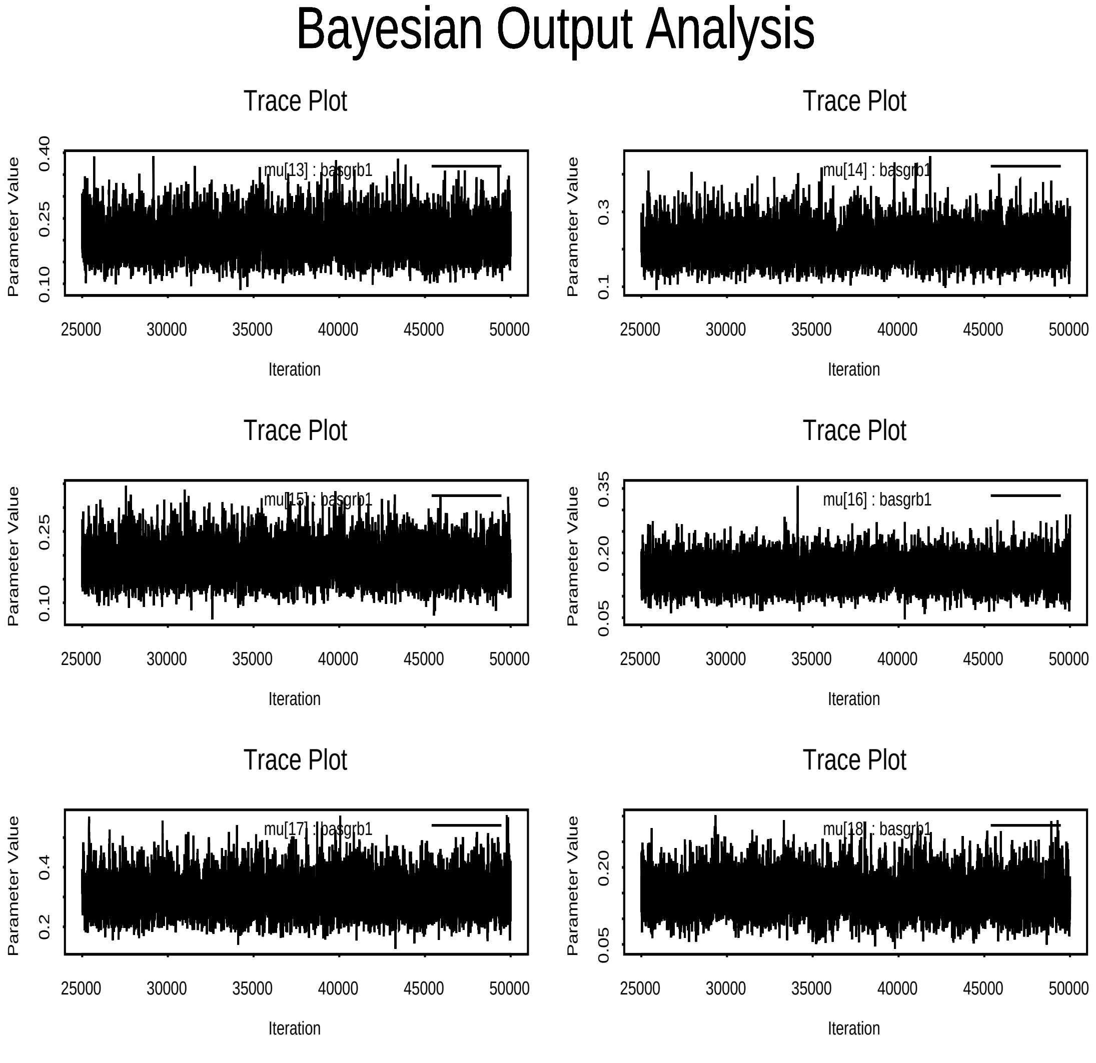


Figure 36: Application study, sampling histories: Selected μ_c parameters, GMIRT Model

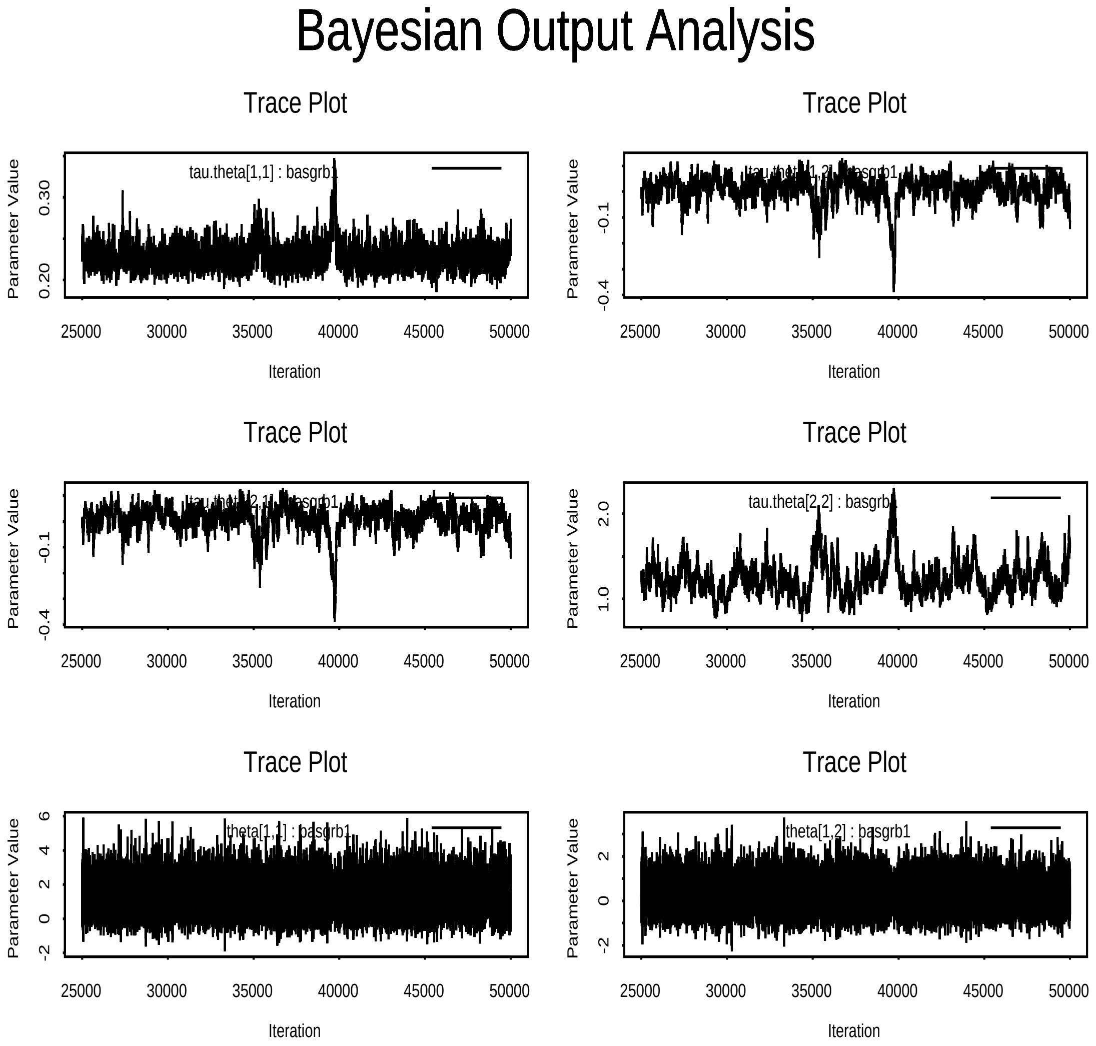


Figure 37: Application Study, sampling histories: Selected τ_θ parameters, GMIRT Model

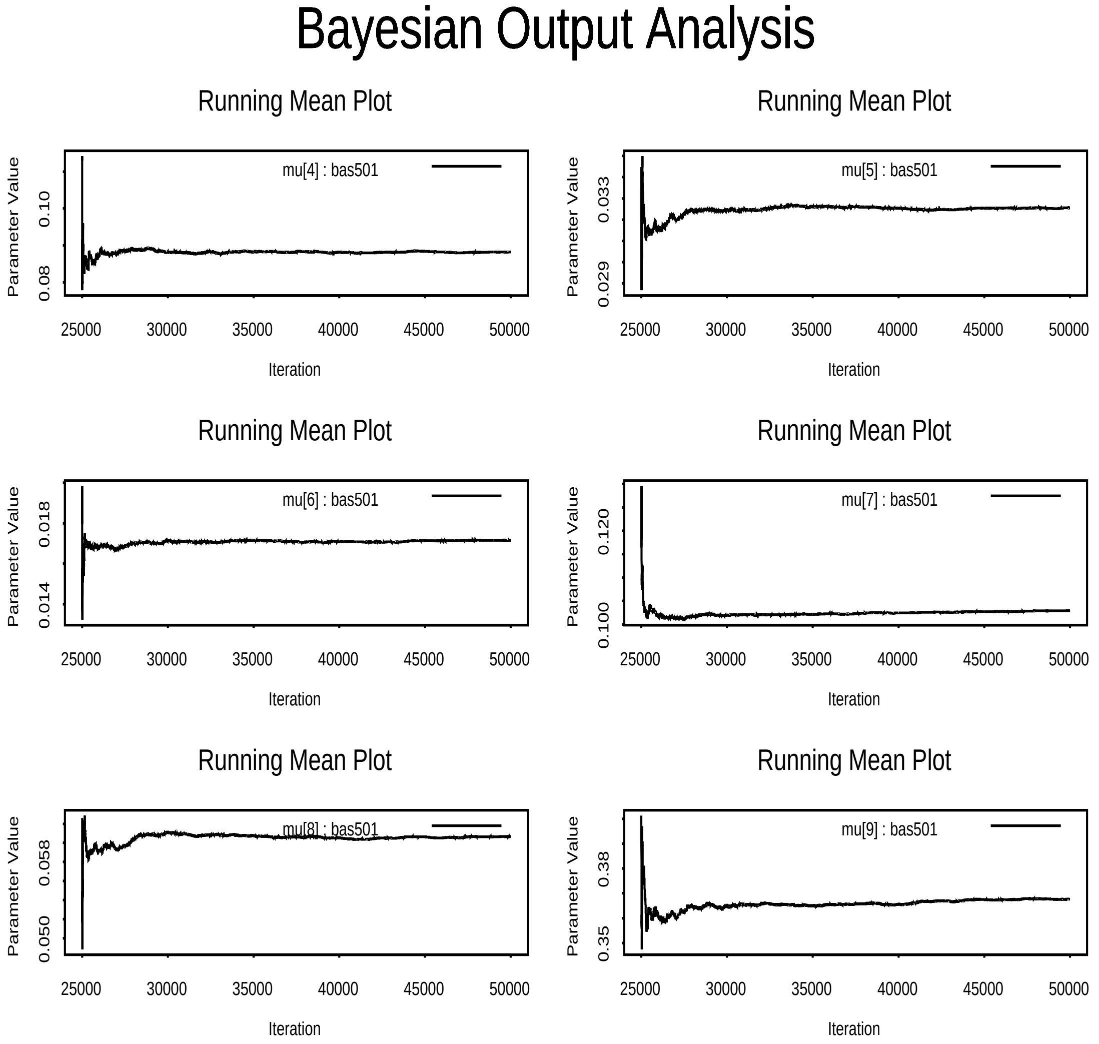


Figure 38: Application study, running mean plot: Selected μ_c parameters, GMIRT Model

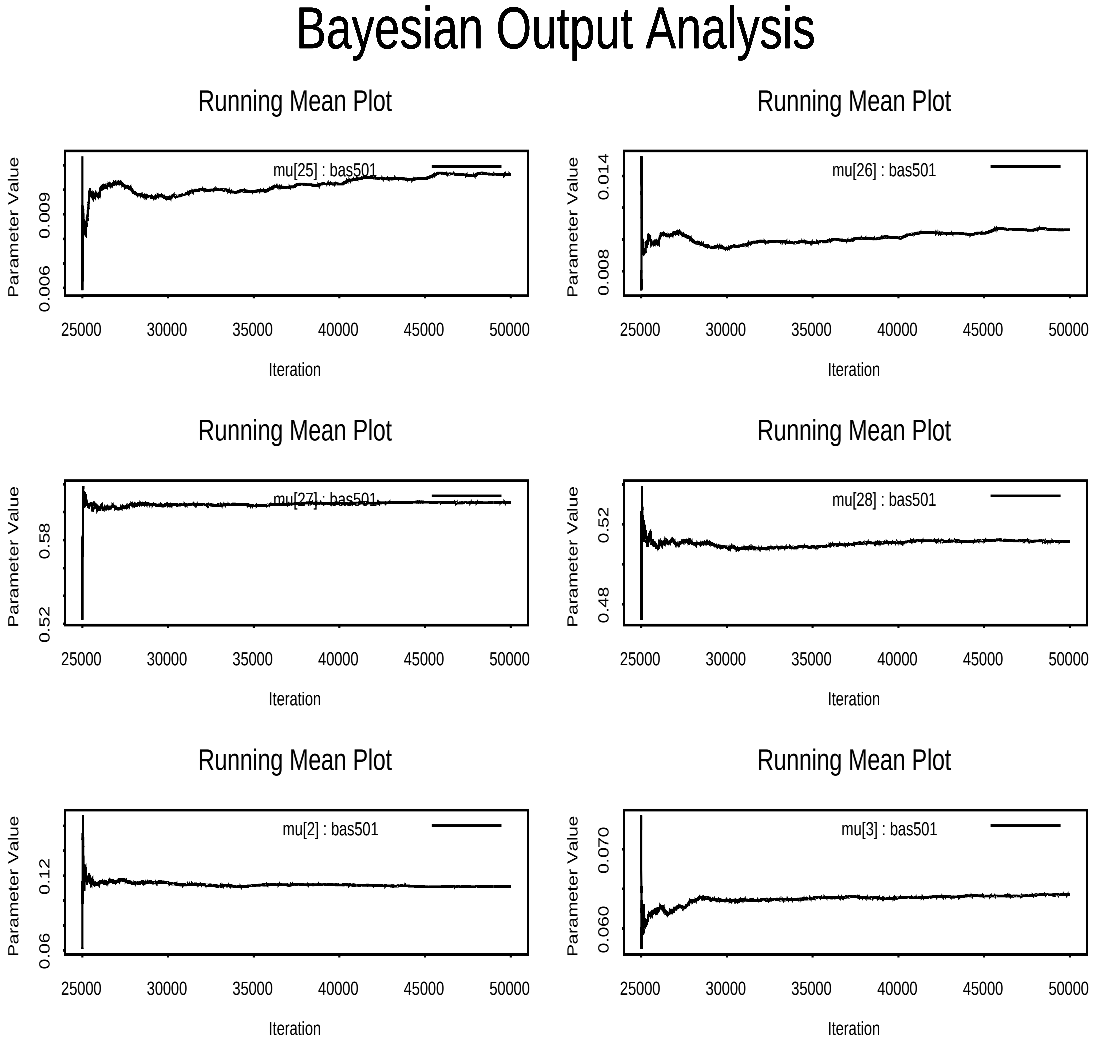


Figure 39: Application study, running mean plot: More selected μ_c parameters, GMIRT Model

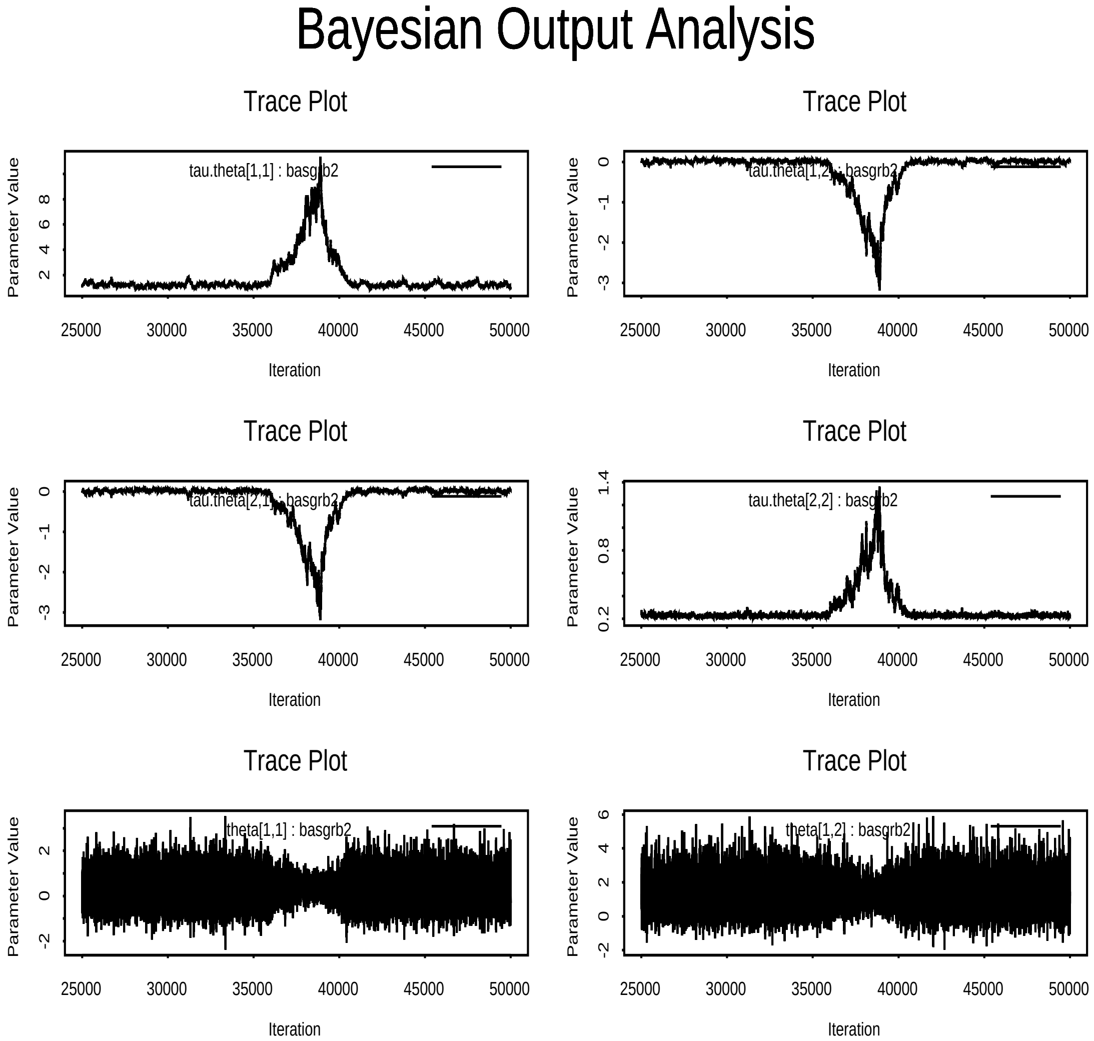


Figure 40: Application study, sampling histories: τ_θ and selected θ parameters, GMIRT Model, Bad Chain

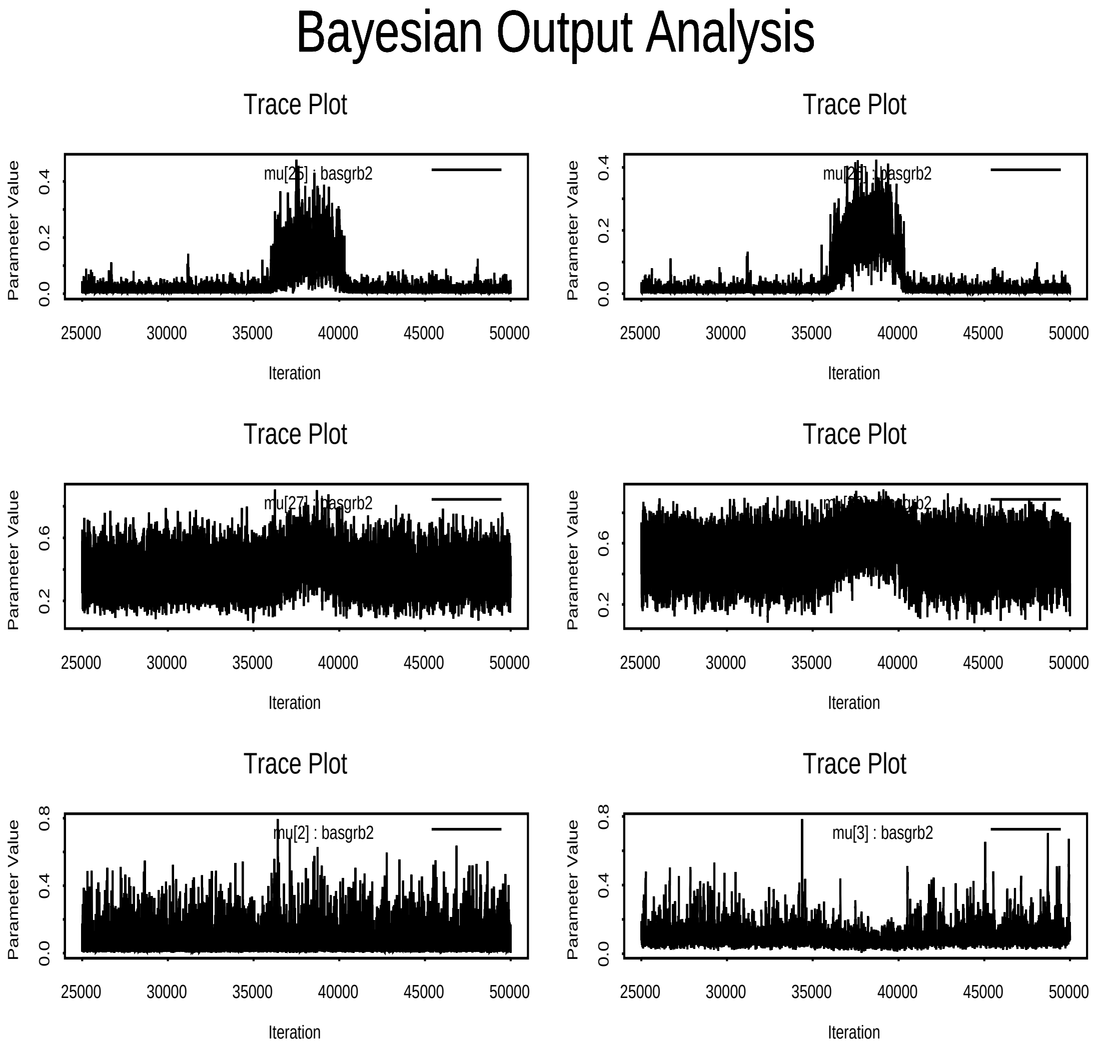


Figure 41: Application study, sampling histories: Selected μ_c Parameters, GMIRT Model, bad chain

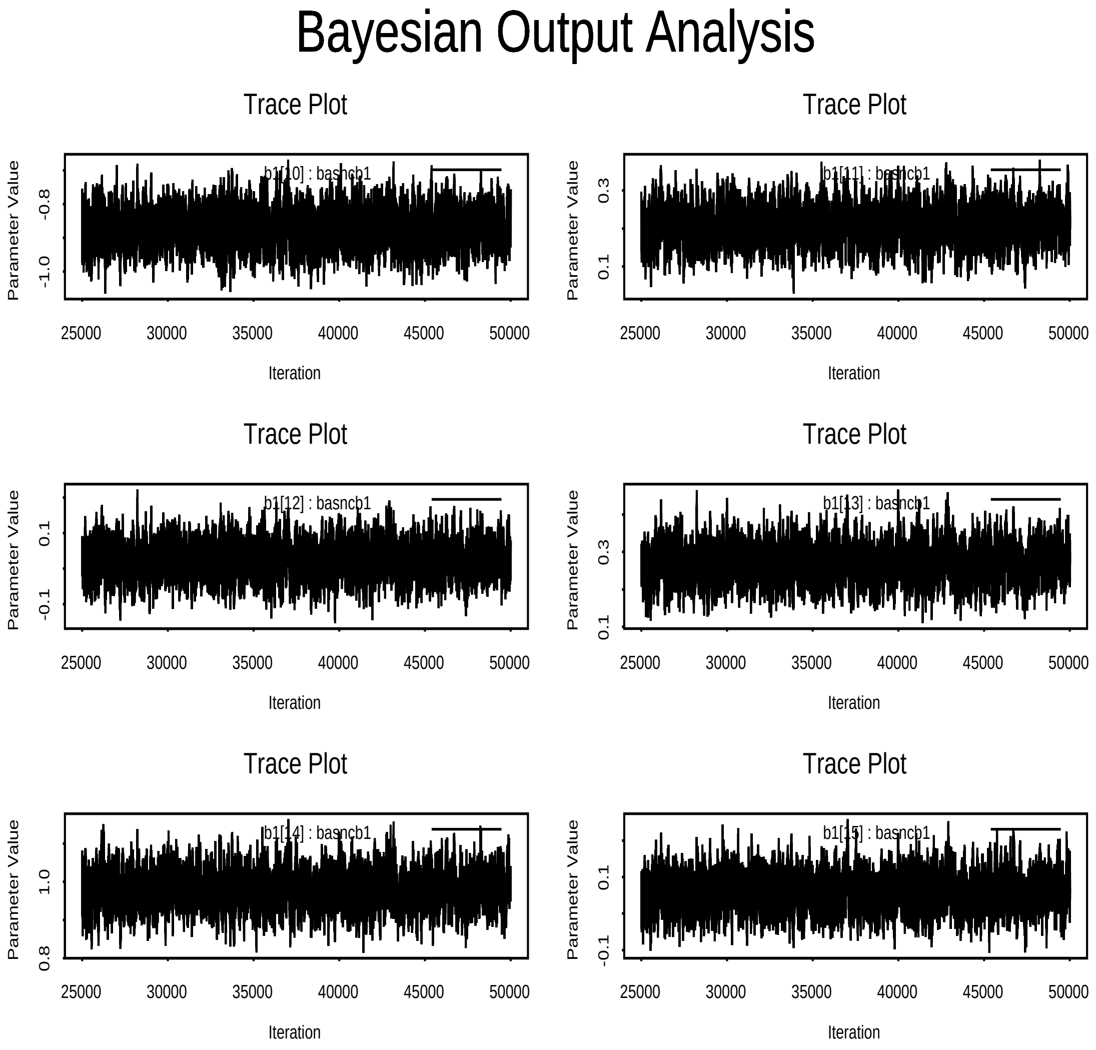


Figure 42: Application study, sampling histories: Selected τ_θ parameters, noncompensatory Model

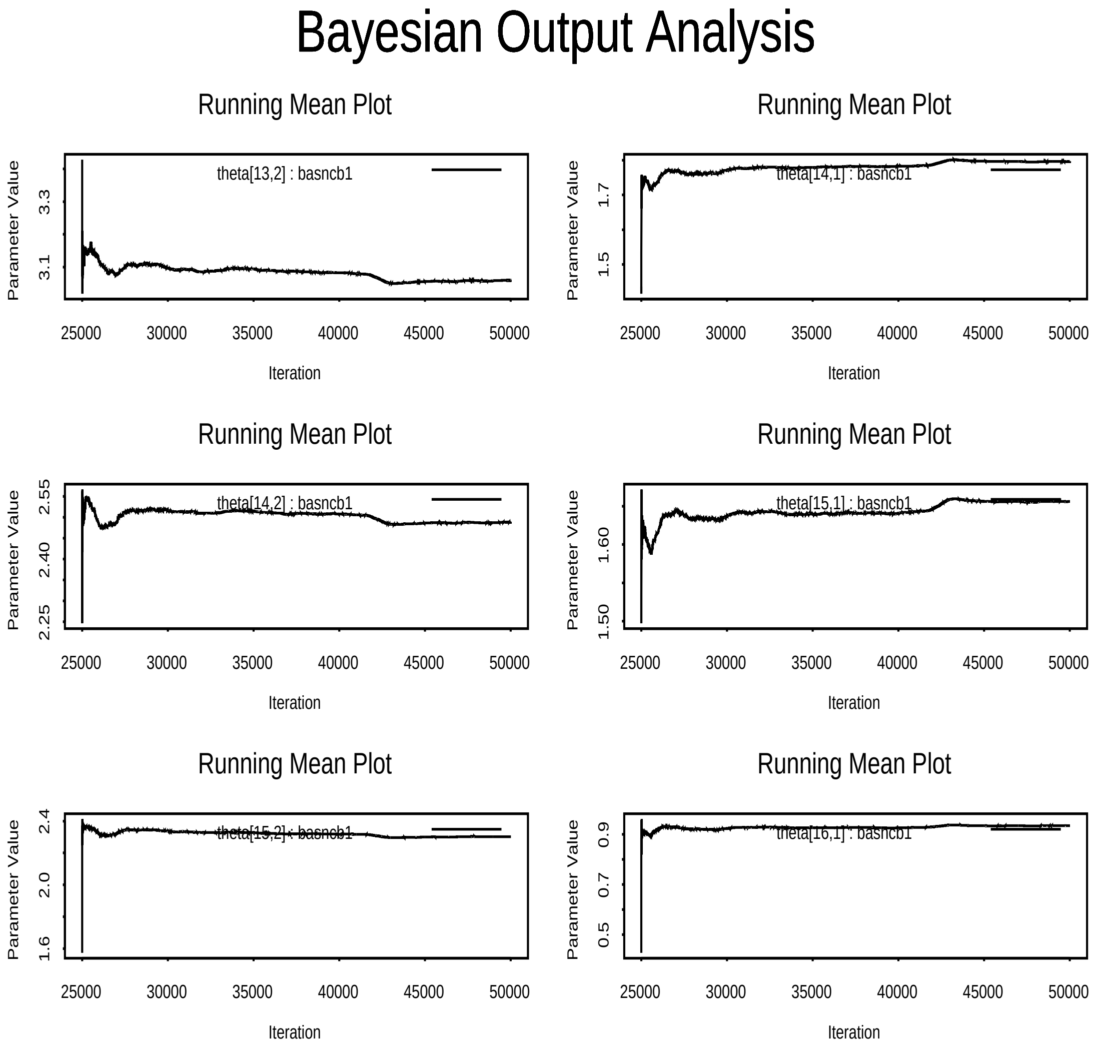


Figure 43: Application study, running mean plot: Selected θ parameters, noncompensatory model

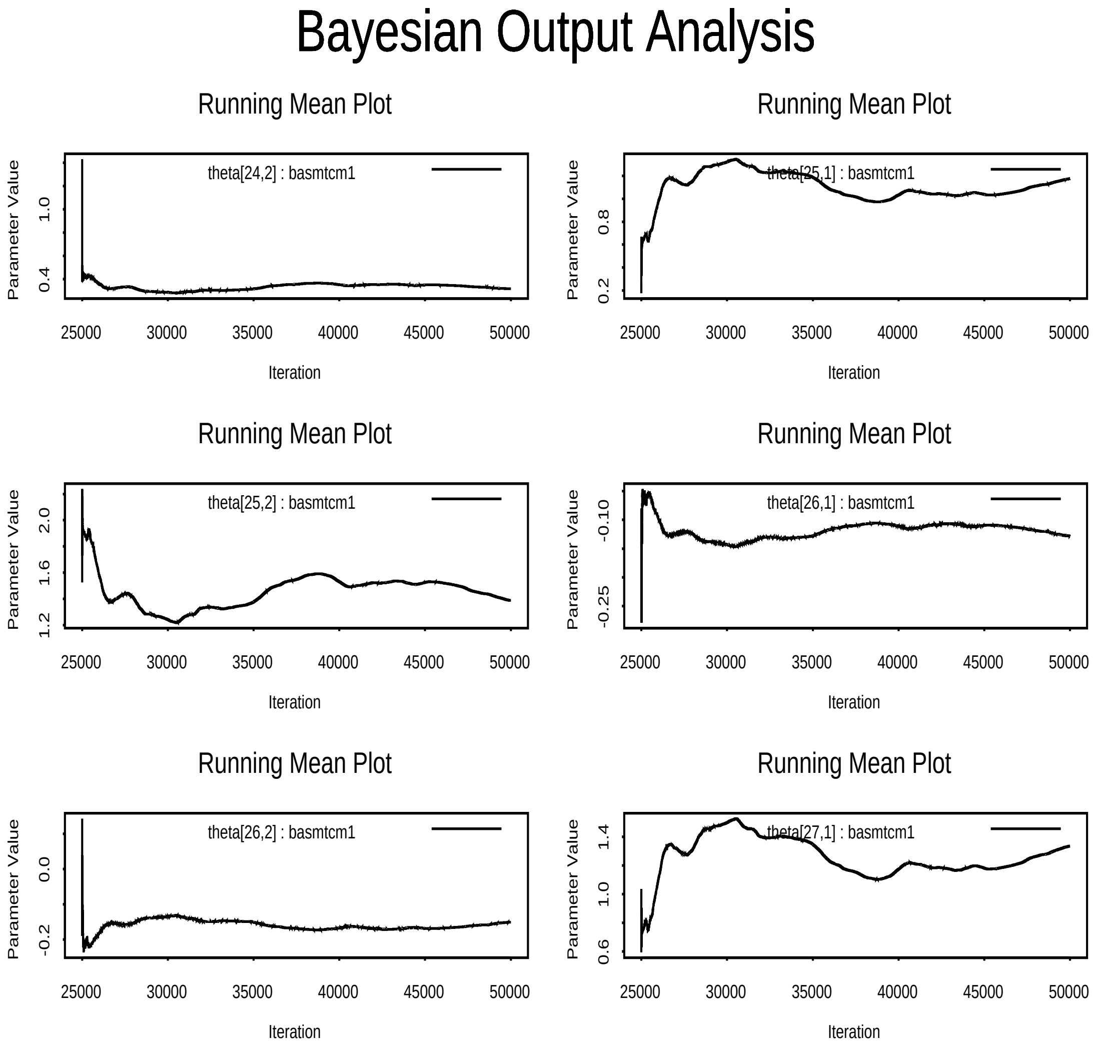


Figure 44: Application study, running mean plot: Selected θ parameters, compensatory model

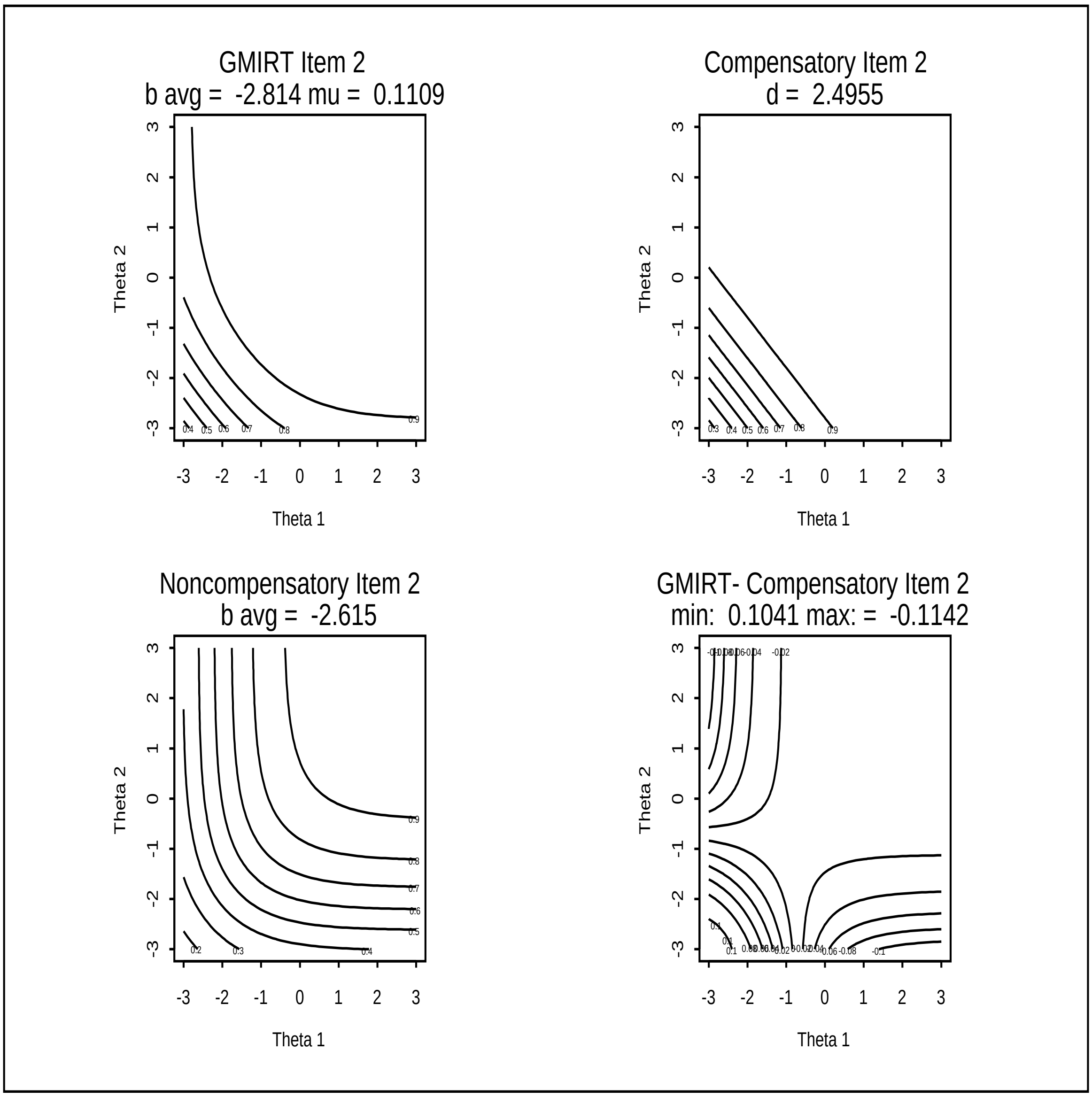


Figure 45: BAS matrices, item 2: Probability contour plots

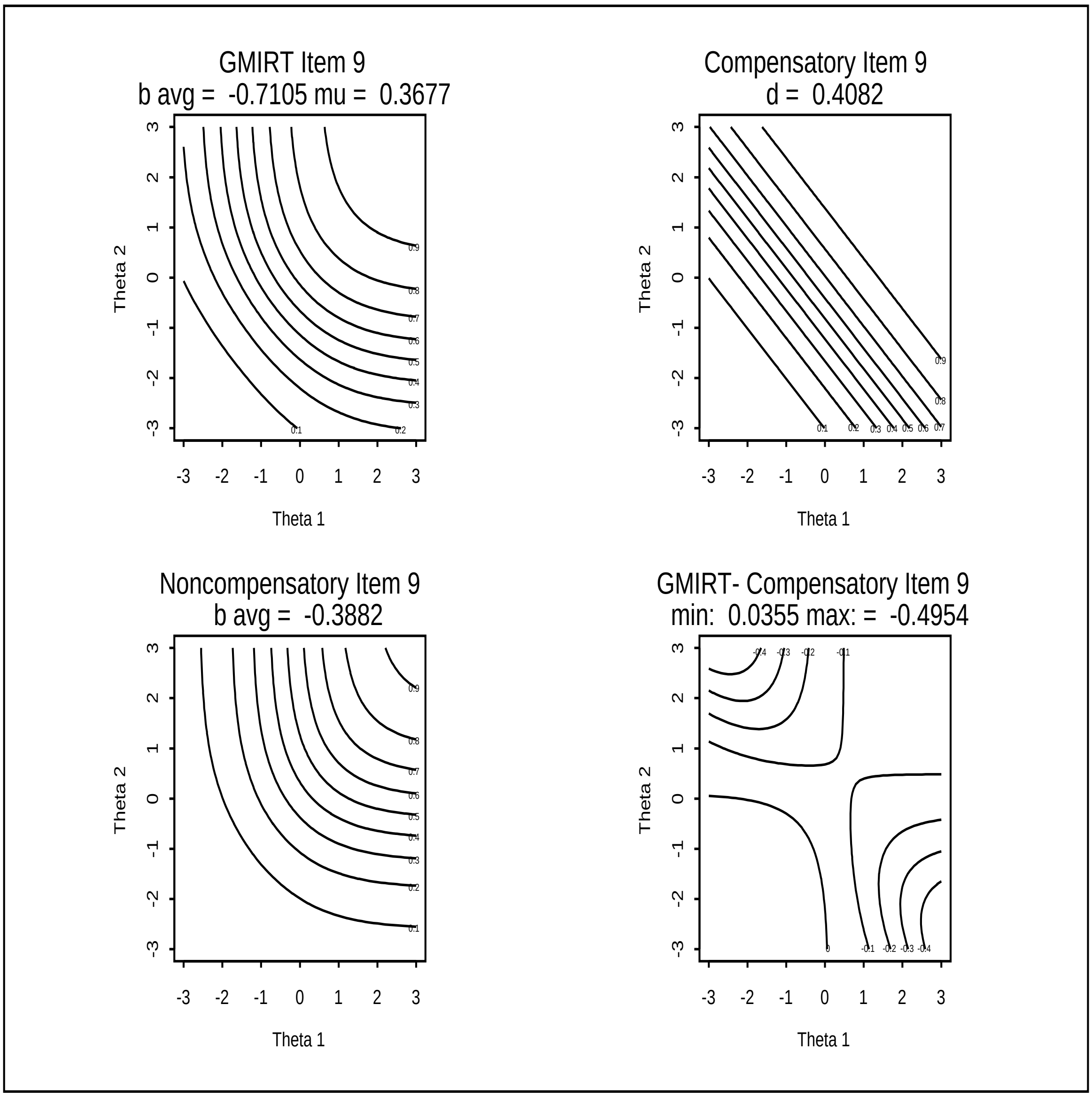


Figure 46: BAS Matrices, item 9: Probability contour plots

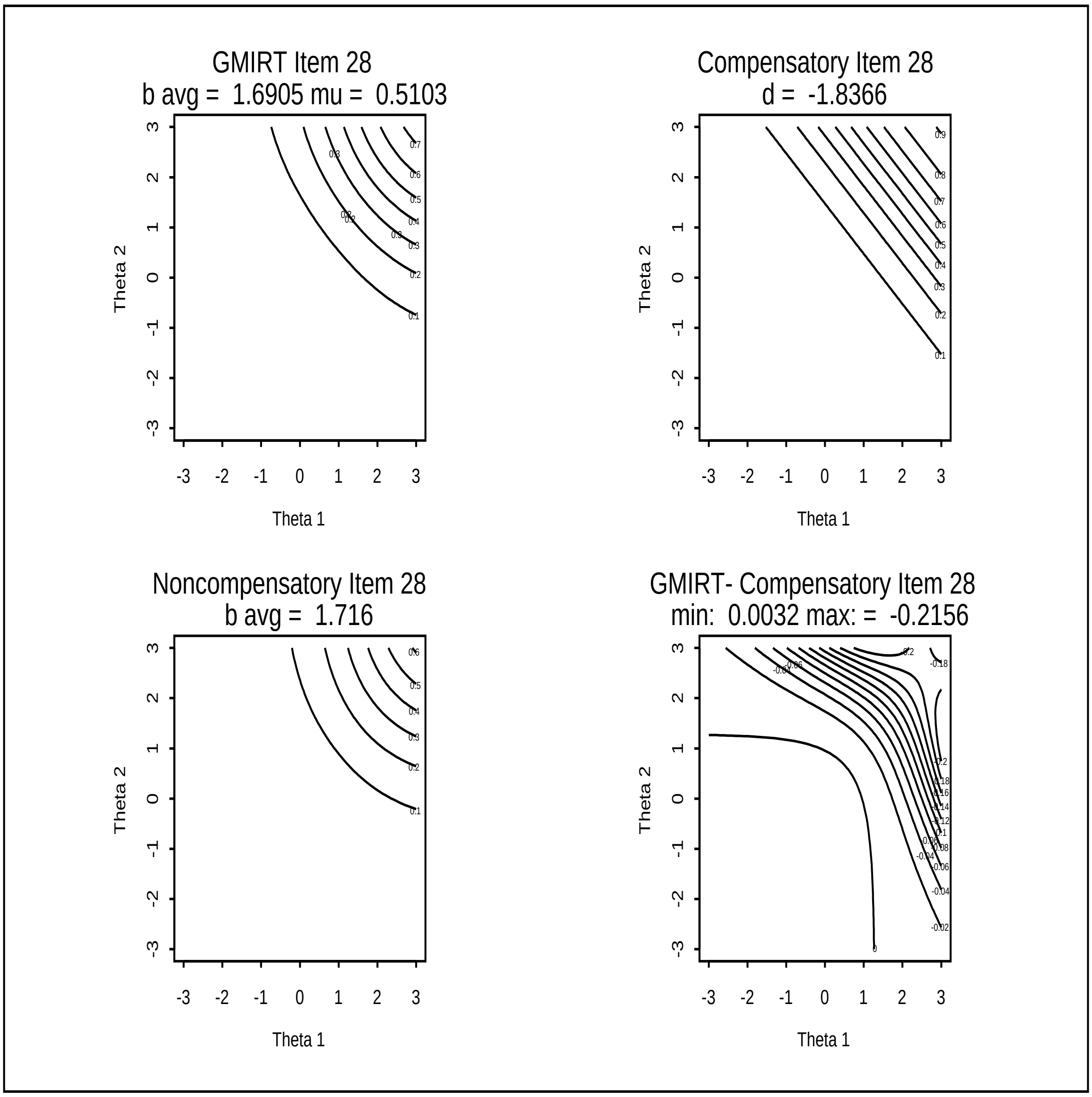


Figure 47: BAS matrices, item 28: Probability contour plots

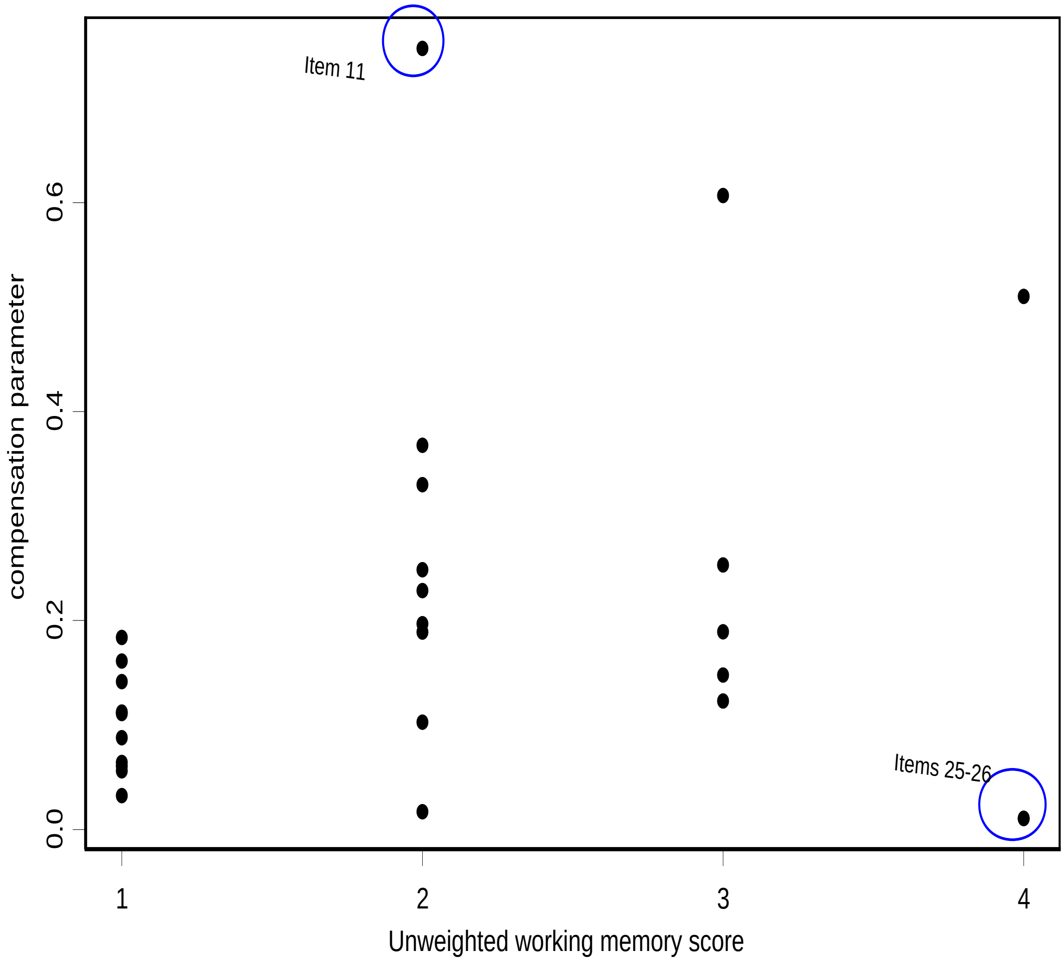


Figure 48: Compensation as function of working memory load, main dataset, unweighted working memory scores

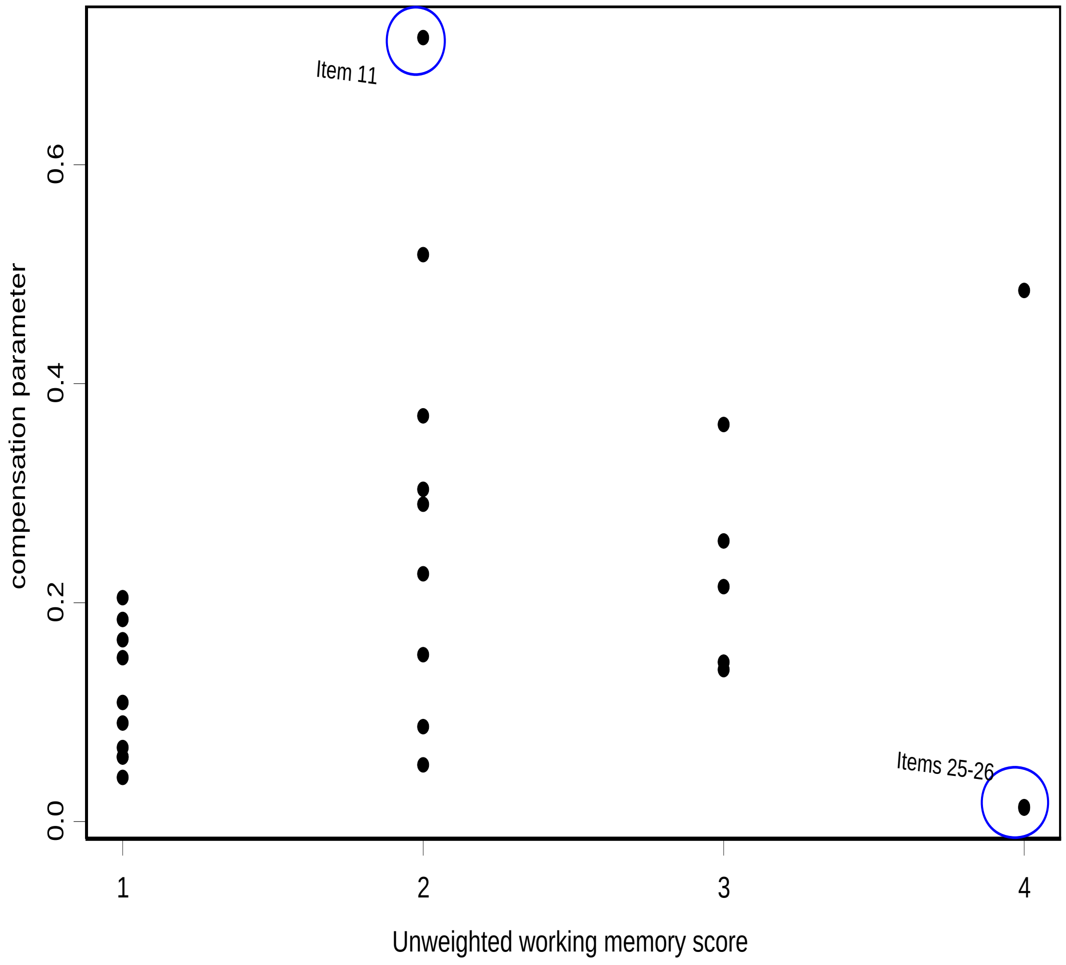


Figure 49: Compensation as function of working memory load, cross-validation dataset, unweighted working memory scores

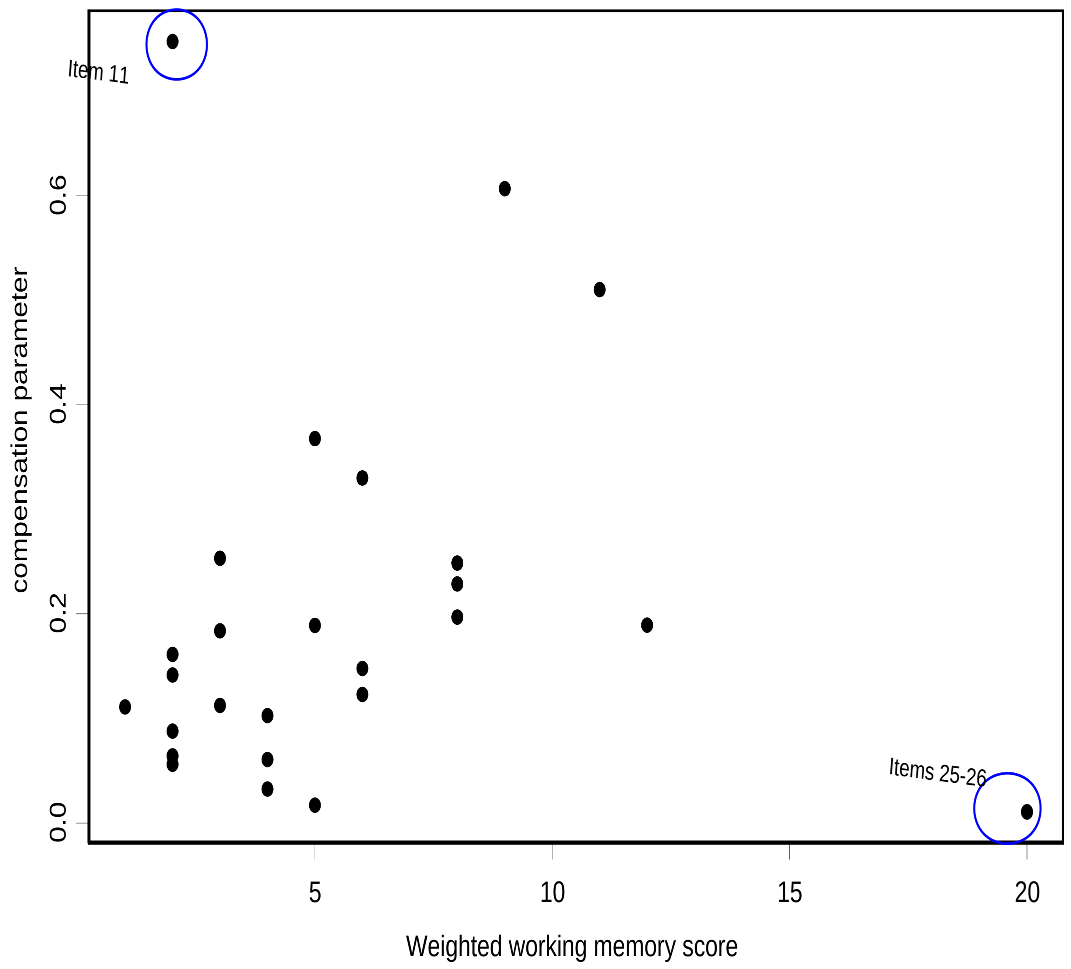


Figure 50: Compensation as function of working memory load, main dataset, weighted working memory scores

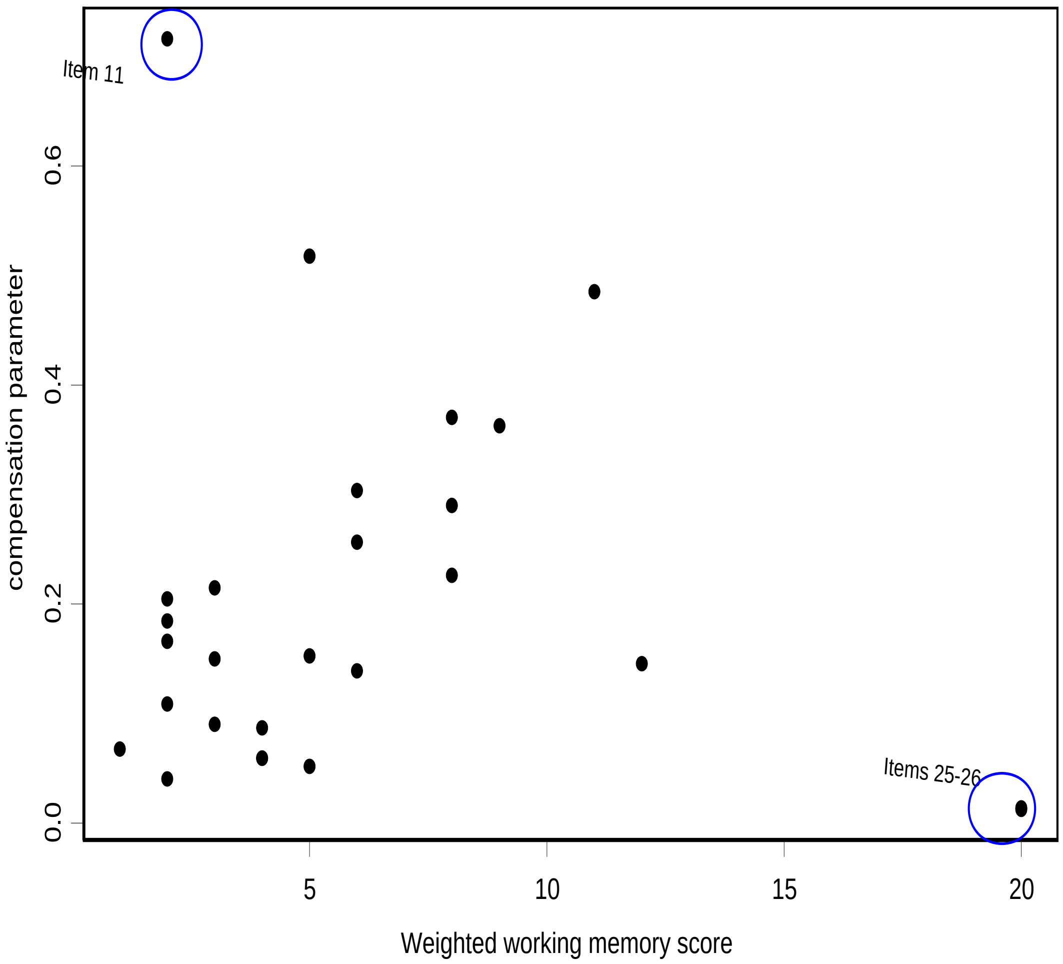


Figure 51: Compensation as function of working memory load, cross-validation dataset, weighted working memory scores

Appendix C. Sample WINBUGS Code

```
### gmirt rasch model, r = 0, 25 items, 6,000 folks ### 03/21/2004

model GRR0256K {
     for (k in 1:I) {
        mu[k] ~dbeta(2,5)
        b1[k] ~ dnorm(0,.5)I(-6,6)
        b2[k] ~ dnorm(0,.5)I(-6,6)
}

      for (j in 1:N) {
         theta[j,1:2] ~dmnorm(mu.theta[1:2],tau.theta[,])
                  I(mint[1:2],maxt[1:2])
}

        tau.theta[1:2,1:2] ~dwish(icovpri[1:2,1:2],2)


     mint[1] <- mu.theta[1]-6
     mint[2] <- mu.theta[2]-6
     maxt[1]<-mu.theta[1]+6
     maxt[2]<-mu.theta[2]+6


     for (j in 1:N) {
       for (k in 1:I) {
        p[j,k] <- (exp(a1[k]*(theta[j,1]-b1[k])+
                     a2[k]*(theta[j,2]-b2[k])))/
            ((1+exp(a1[k]*(theta[j,1]-b1[k])+
                     a2[k]*(theta[j,2]-b2[k])))
                +mu[k]*(exp(a1[k]*(theta[j,1]-b1[k]))
                +exp(a2[k]*(theta[j,2]-b2[k]))))

          r[j,k] ~ dbern(p[j,k])
}}}



#data list(N=6000
I=25,a1=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
         ,1), a2
=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
   1,1,1,1,1,1,1,1,1),
b1=c(0,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
     NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA),
b2=c(0,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
    NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA),
mu.theta=c(0,0),icovpri=structure(.Data=c(1,0,0,1),
    .Dim=c(2,2)), mu=
c(0,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
   NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA),
 r=structure(.Data=c(0,1,0,0,0,1,0,0,1,1,0,0,0,
        1,0,0,0,1,1,1,0,1,0,0,0,
0,0,1,1,0,1,0,0,1,1,1,1,0,1,1,1,0,
            1,1,1,1,0,0,0,1 . . .
1,1,0,0,0,1,1,0,0,0,0,1,0,1,0,1,0,0,0,1,0,1,0,1,0
),.Dim=c(6000,25)))
```