

1-1-1997

Accuracy of parameter estimation in polytomous IRT models.

Chung Park
University of Massachusetts Amherst

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

Recommended Citation

Park, Chung, "Accuracy of parameter estimation in polytomous IRT models." (1997). *Doctoral Dissertations 1896 - February 2014*. 5302.
https://scholarworks.umass.edu/dissertations_1/5302

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

UMASS/AMHERST



312066 0264 0742 4

ACCURACY OF PARAMETER ESTIMATION IN POLYTOMOUS IRT MODELS

A Dissertation Presented

by

CHUNG PARK

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

DOCTOR OF EDUCATION

September 1997

School of Education

© Copyright by Chung Park, 1997

All Rights Reserved

ACCURACY OF PARAMETER ESTIMATION IN POLYTOMOUS IRT MODELS

A Dissertation Presented

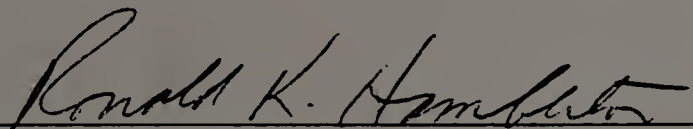
by

CHUNG PARK

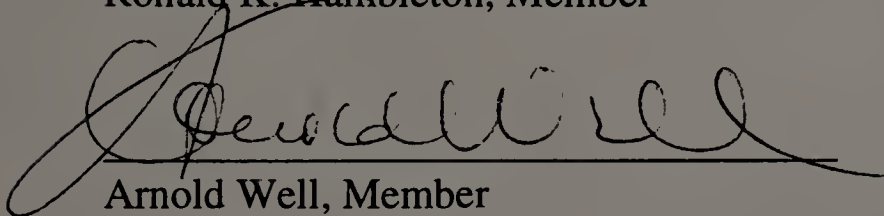
Approved as to style and content by:



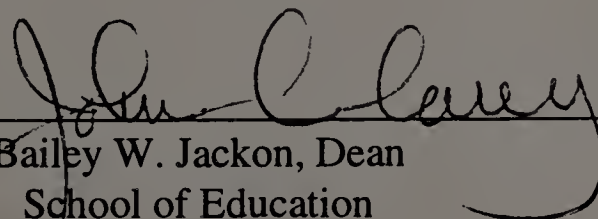
H. Swaminathan, Chair



Ronald K. Hambleton, Member



Arnold Well, Member



Bailey W. Jackson, Dean
School of Education

To my parents and my sister, young
whose lifelong trust and support are beyond the estimation space

ACKNOWLEDGMENTS

There appear to be innumerable faces I should acknowledge as I near the end. Without their help and encouragement, I could not have completed this final step of my doctoral program. Words cannot adequately express my gratitude to them.

I should like to express my appreciation to my committee, Dr. H. Swaminathan, Dr. Ronald K. Hambleton, and Dr. Arnie Well. In particular, I am deeply indebted to Dr. H. Swaminathan for serving as my advisor and chair of my dissertation committee and for providing unlimited support. He has always been immensely generous with his time, his encouragement, and advices. His guidance and suggestions led me to further understanding of my dissertation subject and new insights and questions. His continuous reassurance gave me confidence which I needed during the whole process of this program. He has also devoted his measureless effort to edit my drafts and refine my writing, and I have learned a great deal as a result. More than that, I learned patience and simplicity from him. For this I will always be in his debt.

I am pleased to acknowledge the support of Dr. R. K. Hambleton. He helped me start a graduate student in the program and provided constant encouragement that enabled me with the confidence to undertake every new step in the program. He also provided an outstanding model as both a researcher and teacher through my program of study.

I would like to thank Dr. Arnie Well for serving on this committee. I appreciate the time he has given cheerfully and graciously.

I also expand my appreciation to Dr. Eiji Muraki in Educational Testing Service, who gave me his valuable time and assisted me with the running of the computer program

PARSCALE which was crucial for the completion of my dissertation. He helped me understand the program PARSCALE and the underlying framework of the models in the program.

I have been especially lucky to have Dr. Nancy Allen in Educational Testing Service as my mentor and friend. She gave me opportunities to be involved in a NAEP performance assessment project in the summer of 1993 and taught me a great deal about the process of developing research. She always listened to me and gave me critical but warm comments that helped me to restore perspective at times when I felt perplexed.

I would thank Peg Louraine, who provided assistance and information about all of the procedures and guidelines for completing the program.

Finally, I should like to express appreciation to my family, friends and Professors in Korea. My family has allowed me to be away from home and waited for me with unconditional trust and unlimited support. My friends and Professors have supported me with constant encouragement and deep empathy. Their constant support and reassurance enabled me to study in the U.S.A. in peace and will keep me going.

ABSTRACT

ACCURACY OF PARAMETER ESTIMATION IN POLYTOMOUS IRT MODELS

SEPTEMBER 1997

CHUNG PARK, B. S. , SUNG KYUN KWAN UNIVERSITY, KOREA

M. ED., SEOUL NATIONAL UNIVERSITY, KOREA

ED. D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by : Professor Hariharan Swaminathan

Procedures based on item response theory (IRT) are widely accepted for solving various measurement problems which cannot be solved using classical test theory (CTT) procedures. The desirable features of dichotomous IRT models over CTT are well known and have been documented by Hambleton, Swaminathan, and Rogers (1991). However, dichotomous IRT models are inappropriate for situations where items need to be scored in more than two categories. For example, in performance assessments, most of the scoring rubrics for performance assessment require scoring of examinee's responses in ordered categories. In addition, polytomous IRT models are useful for assessing an examinee's partial knowledge or levels of mastery. However, the successful application of polytomous IRT models to practical situations depends on the availability of reasonable and well-behaved estimates of the parameters of the models. Therefore, in this study, the behavior of estimators of parameters in polytomous IRT models were examined.

In the first study, factors that affected the accuracy, variance, and bias of the marginal maximum likelihood (MML) estimators in the generalized partial credit model

(GPCM) were investigated. Overall, the results of the study showed that the MML estimators of the parameters of the GPCM, as obtained through the computer program, PARSCALE, performed well under various conditions. However, there was considerable bias in the estimates of the category parameters under all conditions investigated. The average bias did not decrease when sample size and test length increased. The bias contributed to large RMSE in the estimation of category parameters. Further studies need to be conducted to study the effect of bias in the estimates of parameters on the estimation of ability, the development of item banks, and on adaptive testing based on polytomous IRT models.

In the second study, the effectiveness of Bayesian procedures for estimating parameters in the GPCM was examined. The results showed that Bayes procedures provided more accurate estimates of parameters with small data sets. Priors on the slope parameters, while having only a modest effect on the accuracy of estimation of slope parameters, had a very positive effect on the accuracy of estimation of the step difficulty parameters.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGMENTS	v
ABSTRACT	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiv
 CHAPTER	
1. INTRODUCTION	1
2. ITEM RESPONSE THEORY MODELS	5
2.1 Item Response Theory	5
2.2 Properties (advantages) of IRT	6
2.3 Dichotomous IRT Models	8
2.4 Polytomous IRT Models	10
2.4.1 Models for Ordinal Responses	11
2.4.1.1 The Graded Response Model	12
2.4.1.2 The Partial Credit Model	18
2.4.1.3 Comparison of the GRM and the PCM	21
2.4.1.4 The Rating Scale Model	23
2.4.1.5 The Generalized Partial Credit Model	26
2.4.2 Model for Nominal Responses	27
2.4.2.1 The Nominal Response Model	27
3. REVIEW OF THE LITERATURE	30
3.1 Estimation Procedures for IRT Models	30
3.2 MMLE Procedure for Polytomous IRT Models	31
3.3 Previous Research on MMLE Procedure	33
3.4 Bayesian Estimation of Parameters in Polytomous IRT Models	38
3.5 Summary	42

4.	DESIGN OF THE STUDY AND METHODOLOGY	44
4.1	Overview of Study	44
4.2	Design of Study	44
	4.2.1 Test Characteristics	45
	4.2.1.1 Test Length	45
	4.2.1.2 The Number of Response Categories	46
	4.2.1.3 Item Parameter Values	46
	4.2.2 Characteristics of the Calibration Sample	47
	4.2.2.1 Sample Size	47
	4.2.2.2 Ability Distribution and the Minimum Number of Examinees in Each Category	48
	4.2.2.3 Estimation Procedure	49
4.3	Data Generation	53
4.4	Criteria for Evaluating Adequacy of the Estimates	55
4.5	Calibration	58
5.	RESULTS	62
5.1	Introduction	62
5.2	Results of Study I	62
	5.2.1 Accuracy of Estimation	62
	5.2.2 Variance and Bias	68
5.3	Results of Study II	72
	5.3.1 Accuracy of Estimation	72
	5.3.2 Variance and Bias	83
	5.3.3 Item Level Analysis on the Accuracy of Estimation	95
6.	SUMMARY AND CONCLUSIONS	114
6.1	Summary and Conclusions for Study I	114
6.2	Summary and Conclusions for Study II	117
6.3	Significance of Study	119
6.4	Delimitations and Directions for Further Research	120

APPENDIX : ADDITIONAL FIGURES	123
REFERENCES	130

LIST OF TABLES

Table	Page
1. Factorial design with 4 factors : 2 x 3 x 4 x 4	51
2. Prior distributions for the slope and the threshold parameters	52
3. True item parameter values for 3 category items	60
4. True item parameter values for 5 category items	61
5. The average RMSE across all conditions for 3 and 5 category items	63
6. Results of ANOVA for Root Mean Squared Error (RMSE)	64
7. The average variance across all conditions for 3 and 5 category items	69
8. Results of ANOVA for variance	70
9. Result of ANOVA for bias	71
10. The average bias across all conditions for 3 and 5 category items	73
11. Average RMSE of estimates of slope parameters across different priors for 3 and 5 category items	76
12. Average RMSE of estimates of step difficulty parameters across different priors for 3 and 5 category items	77
13. Results of ANOVA for RMSE	78
14. Average variance of estimates of slope parameters across different priors for 3 and 5 category items	84
15. Average variance of estimates of step difficulty parameters across different priors for 3 and 5 category items	85
16. Results of ANOVA for variance	86
17. Average bias of estimates of slope parameters across different priors for 3 and 5 category items	89

18.	Average bias of estimates of step difficulty parameters across different priors for 3 and 5 category items	90
19.	Results of ANOVA for Bias	91

LIST OF FIGURES

Figure	Page
1. Boundary characteristic curves for four category item	14
2. ICCCs with four ordinal responses under the GRM and the PCM	16
3. Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on normal distribution	65
4. Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on normal distribution	66
5. Average bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items	74
6. Average bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items	75
7. Average RMSE of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items	80
8. Average RMSE of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items	81
9. Average variance of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items	87
10. Average variance of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items	88
11. Average bias of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items	93
12. Average bias of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items	94
13. RMSE of estimates of slope parameters for each item in 3 category 9 items ...	98
14. RMSE of estimates of the first step difficulty parameters for each item in 3 category 9 items	99

15.	RMSE of estimates of the second step difficulty parameters for each item in 3 category 9 items	100
16.	RMSE of estimates of slope parameters for each item in 3 category 18 items	101
17.	RMSE of estimates of the first step difficulty parameters for each item in 3 category 18 items	102
18.	RMSE of estimates of the second step difficulty parameters for each item in 3 category 18 items	103
19.	RMSE of estimates of slope parameters for each item in 5 category 9 items	104
20.	RMSE of estimates of the first step difficulty parameters for each item in 5 category 9 items	105
21.	RMSE of estimates of the second step difficulty parameters for each item in 5 category 9 items	106
22.	RMSE of estimates of the third step difficulty parameters for each item in 5 category 9 items	107
23.	RMSE of estimates of the fourth step difficulty parameters for each item in 5 category 9 items	108
24.	RMSE of estimates of the slope parameters for each item in 5 category 18 items	109
25.	RMSE of estimates of the first step difficulty parameters for each item in 5 category 18 items	110
26.	RMSE of estimates of the second step difficulty parameters for each item in 5 category 18 items	111
27.	RMSE of estimates of the third step difficulty parameters for each item in 5 category 18 items	112
28.	RMSE of estimates of the fourth step difficulty parameters for each item in 5 category 18 items	113

A.1	Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on uniform distribution	124
A.2	Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on uniform distribution	125
A.3	Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on positively skewed distribution	126
A.4	Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on positively skewed distribution	127
A.5	Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on negatively skewed distribution	128
A.6	Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on negatively skewed distribution	129

CHAPTER 1

INTRODUCTION

Procedures based on item response theory (IRT) are widely accepted for solving various measurement problems which can not be solved when using classical test theory (CTT) procedures. The advantages of IRT over classical test theory include: (1) item parameters that are independent of the subpopulations of examinees to which an instrument (or a test) is administered and (2) ability parameters that are independent of the items used. An important distinction between IRT and CTT is that IRT is item oriented while CTT is test oriented. This feature permits assembling items so that a test with desired characteristics can be constructed. A further advantage of IRT is that a measure of precision for each level ability score is available more readily than with CTT (Hambleton, Swaminathan, and Rogers, 1991).

Item response theory (IRT) models that can deal with dichotomous scored items are well-developed and are now in common use. However, dichotomous IRT models restrict scoring of examinee responses to "right" or "wrong". To use dichotomous IRT models for data that have multiple response categories in items, some category responses have to be collapsed into two categories, e.g., in rating scales, responses for category 1, 2, and 3 could be recorded as "0" and 4 and 5 are recorded as "1" ; for multiple-choice items all incorrect option responses are recorded as "0" while correct response is recorded as "1". Through those procedures, some information in multicategory responses will be lost (Carlson, 1996). While IRT models that permit polytomous scoring (nominal as well as

ordinal) were introduced in the sixties and early seventies (Samejima, 1969; Bock, 1972), they have not received wide attention until recently.

Currently, polytomous IRT models are receiving an increasing attention with emphasis on performance assessment, and are emerging as the models of choice for the analysis of the type of the data obtained when the response of an examinee to an item is scored on a scale rather than as right/wrong. Such models make it possible to assess an examinee's partial knowledge (as in performance assessment, for example) and to analyze rating scales of the Likert type. Development of polytomous response models allows the information from those multicategory responses to be used. With the advent of computer programs for estimating parameters, applications of those models have begun to flourish. Their applications to a variety of situations have been documented by several researchers (see for example, Carlson, 1996; Dodd, DeAyala & Koch, 1995; and Potenza & Dorans, 1995).

There are two types of polytomous IRT models. One is appropriate for items that have the response categories arranged in the order of attainment or intensity. The ordered response models can be applied in variety of situations such as grading essay items, attitude measurement, assessment of partial knowledge, and assessment of proficiency attainment as in performance assessment. The other is appropriate for when the response categories are nominal in nature. These models are appropriate for comparison of examinees at various ability levels on their choices of the distractors for diagnostic purposes .

Realizing the promise that polytomous IRT models hold for assessment is predicated on accurate estimation of parameters in these models. A few studies (Choi,

Cook, & Dodd, 1996; De Ayala, 1995; Reise & Yu, 1990; and Walker-Bartnick, 1990) have focused on the problem of estimation of parameter estimates in polytomous IRT models. These studies have provided useful information regarding the effects of certain factors on parameter estimation. Many issues, however, remain to be addressed with respect to the problem of estimation in polytomous IRT models. For example, the effects of interaction among such factors as test length, number of examinees, the number of response categories, ability distributions, and the particular polytomous model on the estimation of parameters are not known. Only through a systematic study of the factors can recommendations be made to practitioners about the data requirements for satisfactory estimation of the parameters of polytomous IRT models.

The primary purpose of this study is therefore to study marginal maximum likelihood (MML) and Bayesian estimation procedures for estimation of parameters in polytomous IRT models as implemented by the available computer program. MML procedure is the commonly implemented procedure to obtain estimates of parameters in polytomous IRT models. Bayesian approach is an alternative to solve the problems which may be occurred when MML procedure is applied. This dissertation focused on examining properties of estimators in one of ordinal polytomous IRT models, the generalized partial credit model (GPCM). The effects of factors that affecting parameter estimation in the GPCM were examined systematically for the purpose of making recommendation regarding data requirements for parameter estimation in polytomous IRT models.

In the first study, the behavior of estimators in the GPCM and the effects of various factors such as sample size, test length, the number of categories in each item,

and ability distribution on them were examined. More specifically, the properties of the MML estimators such as accuracy, bias, and consistency for the GPCM were examined under various conditions. In the second study, the effectiveness of a Bayesian approach to estimation in the GPCM was investigated and compared the Bayesian procedure with the marginal maximum likelihood procedure. In particular, the issues investigated were (a) the effects of specifying priors on the item parameters and (b) the accuracy, variance, and bias of estimators.

This dissertation consists of six chapters. Chapter 2 explains IRT models including polytomous IRT models. Chapter 3 contains the review of the literature of estimation procedures for polytomous IRT models. Chapter 4 describes the design of the study and methodology. Chapter 5 presents the results of the study. The final chapter draws summary and conclusions from the study.

CHAPTER 2

ITEM RESPONSE THEORY MODELS

2.1 Item Response Theory

Item response theory models specify the relationship between observable examinee item performance and the unobservable trait or ability assumed to underlie performance on the test (Hambleton & Swaminathan, 1985, p. 9). The relationship is expressed in the form of a mathematical function. The function (item response models) is based on the assumptions one is willing to make about the item set under investigation. While the item response models differ from one another in the specific mathematical function, they have a common assumption, that of unidimensionality of the trait.

The fundamental assumption of (unidimensional) item response models is that a test measures a single latent trait or ability, i.e., the assumption of unidimensionality. While theoretically IRT models can be formulated for multidimensional traits (Embretson, 1984; Mckinley & Reckase, 1982; Samejima, 1974), those models are not well developed and will not be addressed in this study.

Equivalent to the assumption of dimensionality is the local independence. When the complete latent space (unidimensional in this case) is specified, the item responses are independent of one another when the ability level is fixed at a value. An important consequence of this assumption is that for an examinee, the probability of an observed response pattern is the product of the probabilities of the observed responses on the

individual items. This result is of fundamental importance in the estimation of item parameters in IRT.

2.2 Properties (advantages) of IRT

Once the assumptions of IRT model are met (i.e., a set of test items being analyzed fits an unidimensional item response model) and the probability of a correct response follows the specified mathematical function, the property of invariant item and ability parameters holds. That is, item parameters are independent of the subpopulation of examinees for whom the test was designed and the examinee ability is independent of the particular choice of test items used from the set of items. The invariance property has important applications in test development, and trait estimation and sets IRT apart from CTT.

The most important property of unidimensional item response models is that an examinee's ability can be estimated and placed on a common scale with other examinees who are administered different sets of items chosen from a domain of items that have been fitted to the model. This property makes possible adaptive testing where items that are "optimal" for each examinee can be selected for administration. The advantage of such a testing scheme is that tests no longer need to encompass a wide range of difficulty to ensure adequate accuracy measurement throughout the ability continuum.

IRT facilitates construction and maintenance of item banks. Sets of items can be calibrated independently using different samples of examinees and then be combined to form an item bank with all item 'statistics' on the same scale. When tests are constructed from precalibrated items, the relationship between item parameters and test scores are

known and therefore the tests can be considered "equated". This procedure is referred to as pre-equating, since the tests are placed on a common scale prior to the actual administration. Two tests given to subgroups of the same population can be equated after administration by adapting one of several equating designs of which the most popular design is where common items is embedded in the tests to be equated. Since the item parameters are invariant over subpopulation of examinees, the relationship between the item parameters of the common items in the two tests is established. In turn, this establishes the relationship between ability scores for the two tests, and the need to equate tests in the classical sense is obviated (Lord, 1980, p.205).

The item response models also provide the concepts of item and test information functions. These concepts provide procedures for the assessment of precision and hence are invaluable aids for test construction and item selection. Birnbaum (1968) defined information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability. The general theory of maximum likelihood estimation indicates that the standard error of the estimate of ability is given as the reciprocal of the square root of information. Item information functions provide independent contributions to test information and therefore can be summed to produce in a test information functions. The test information permits a test constructor to select items that together can provide the level of accuracy desired in particular regions of the ability scale. This is of particular importance when tests are constructed with particular purposes in mind such as for selection or placement of candidates.

Once a large item bank has been constructed, the design and construction of equivalent forms of a test or tests for different purposes can be accomplished readily. To

achieve this end one must first determine the desired test information function, which is called target information function. Items are then selected for inclusion in the test until the actual test information function yields a satisfactory approximation to the target test information function (Lord, 1977). The development of equivalent forms with classical testing approaches is not as easy because parallel tests are difficult to construct and the contribution of individual items to the test reliability is difficult to determine in advance.

The invariance property of item response models also provides for the study of item bias (or differential item functioning). Because the item parameters of a set of items, measuring a single dimension must be the same for all subgroups of examinees (Lord, 1980, p.217), when a difference in parameters for an item across subgroups occurs, it must be concluded that the item is differentially functioning across groups.

2.3 Dichotomous IRT Models

The function that expresses the relationship among the trait or ability, θ , the parameters that characterize an item, and the probability, $P_j(\theta)$, of a correct response to an item j is called as the item characteristic curve (ICC) or item response function (IRF) for item j . The curves differ from one another by the number of parameters each model uses to define the shape of ICC. The most popular models employ either one, two, or three item parameters in their respective functions. Lord (1952) proposed an item response model in which the ICC took the form of the normal ogive;

$$P(x_j=1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j) \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \quad j=1,2,\dots,n \quad (1)$$

In this model:

x_j is the response to the j -th item of an n -item test ($x_j = 0, 1$),

$P_j(\theta)$ is the probability of a response of 1 given θ, a_j, b_j, c_j ,

θ is the trait variable or ability

a_j is the discrimination parameter of the j -th item,

b_j is the difficulty or location parameter of the j -th item,

c_j is the lower asymptote of the response function for the j -th item (guessing parameter or pseudo-chance level parameter of j -th item)

Although there can be many item response models based on the mathematical form taken by the item characteristics, the commonly used IRT models involves the logistic distribution function (Birnbaum, 1968) because of computational convenience. The logistic item response model in which the item characteristic curve takes the form of the logistic distribution is,

$$P(x_j=1 | \theta, a_j, b_j, c_j) = c_j + (1 - c_j) \frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}}, \quad j=1, 2, \dots, n \quad (2)$$

The factor 1.7 ensures a close agreement between the logistic response function and one based on the normal ogive.

From an IRT perspective, items may be characterized as differing from one another with respect to difficulty (b_j), the capacity to discriminate (a_j), and guessing or chance-level parameter (c_j). In the 3-parameter logistic model, items differ from one

another with respect to all three item parameters. The 2-parameter logistic model permits items to differ from one another in terms of item discrimination and difficulty parameters, but not guessing parameter, whereas in the one-parameter logistic model only the item difficulty parameter is free to vary (i.e., a_j is assumed to be 1 and c_j is 0). The one parameter logistic model is called the Rasch model (Hambleton & Swaminathan, 1985).

While IRT models offer numerous advantages over classical test theory, IRT models have been restricted for the analysis of dichotomously scored items. To use information from polytomously scored items on which partial credit can be earned, attitude scale items, and personality test items, models that handle multi-category responses are needed.

2.4 Polytomous IRT Models

A series of models for use with multcategory items have been developed by in the field, notably by Andersen (1973), Andrich (1978), Bock (1972), Masters (1982), Muraki (1992), Samejima (1969, 1972), Thissen & Steinberg's (1984). These models may be classified into two categories, ordinal or nominal response models.

Models for ordinal responses are used when the response to an item can be classified into a certain limited number of categories arranged in the order of attainment or intensity. This occurs, for example, when the response to an item can be evaluated according to its degree of attainment of problem solution in the measurement of ability (i.e., in a performance item that requires partial credit) or its degree of intensity of preference to the statement in the measurement of attitude as in a Likert-type statement.

In contrast to models for ordinal responses, models for nominal responses assume that the response to an item is measured at a nominal level of measurement (i.e., unordered responses). Nominal response models are appropriate for studying distractor functioning since the case of multiple-choice items, incorrect alternatives do not represent partially correct answers. However all of these polytomous models are based on assumptions that the item responses depend on a single continuous latent variable and are assumed to be independent, conditional on the value of a latent continuous variable θ .

2.4.1 Models for Ordinal Responses

For ordinal responses, the graded response model (GRM) by Samejima (1969), the rating scale model (RSM) by Andrich (1978), the partial credit model (PCM) by Masters (1982), and the generalized partial credit model (GPCM) by Muraki (1992) are commonly used. The GRM is as an extension to the polytomous case of the two-parameter logistic model for dichotomously-scored items, while the PCM is as an extension to the polytomous case of the one-parameter logistic model or Rasch model. However, the notable distinction between the GRM and the PCM is not the number of parameters but the difference between operating characteristic functions used in those models. The operating characteristic function expresses how the probability of a specific categorical response is formulated according to the law of probability, as well as psychological assumptions about item response behavior. In this section, the GRM and the PCM will be described and compared. The RSM as a special case of the PCM will be described and the GPCM as an extension of the PCM.

2.4.1.1 The Graded Response Model

Samejima (1969) extended the Thurstone's method of successive intervals for dichotomous scored items to more than two, ordered categorized items and introduced a graded response model (GRM). The GRM is appropriate when an examinee's response to an item needs to be scored on the basis of partial correctness (for example, incorrect, partially correct, correct) as in a performance item or on the basis of varying degrees of agreement with the attitude statement as in a Likert-type item. Samejima (1969) categorized the GRM into homogeneous and heterogeneous cases. The homogeneous GRM is for the items in which the thinking process used in solving a given item is assumed to be homogeneous through the whole process, while the heterogeneous GRM assumes that the process consists of different subprocesses. In this paper only the homogeneous case will be handled. That is, in the model the discriminating power should be almost constant throughout the whole thinking process required in solving the problem.

Samejima (1969) showed that the GRM can be reduced to a two-parameter IRT models. The difference between a dichotomous IRT model and the GRM is the number of thresholds that are values of the item variable differentiating response categories. In dichotomous IRT models there is one threshold value, while there may be two or more threshold values in the GRM. The threshold value is called the response category boundary in the GRM, while it is called the item difficulty in a dichotomous IRT model.

Samejima (1969) assumed that any response to an item can be classified into $(m+1)$ ordered categories, scored $k=0, 1, \dots, m$, so that lower-numbered category score represents less of the latent trait measured by the item than do higher-numbered category score. She developed a two-stage process to obtain the probability that an examinee would receive a given category score on an item.

In the first stage the probability that an examinee with ability level θ will receive a given category score k or a higher category score on item j is given by equation (3)

$$P[x \geq k | \theta] = P_{jk}^*(\theta) = \frac{e^{Da_j(\theta - b_{jk})}}{1 + e^{Da_j(\theta - b_{jk})}} \quad (3)$$

where D is the scaling constant 1.7 which maximizes the similarity of the cumulative logistic function to the normal ogive function, b_{jk} is the boundary parameter associated with category score k in item j , a_j is the discrimination parameter of item j , and θ is the ability level; P_{jk}^* is called the boundary (category) characteristic curve. Since the responses to an item j are classified into $m+1$ categories, there are m category boundaries. If the category score k is zero, the probability of responding in category 0 or higher equals 1.0; $P_{j0}^* = 1.0$.

The graphic representations of the functions obtained from equation 3 for a given item can be described as a set of category characteristic curves. Figure 1 depicts a set of category characteristic curves for an item with four categories. While there are four response categories in the item, there are only three boundary curves.

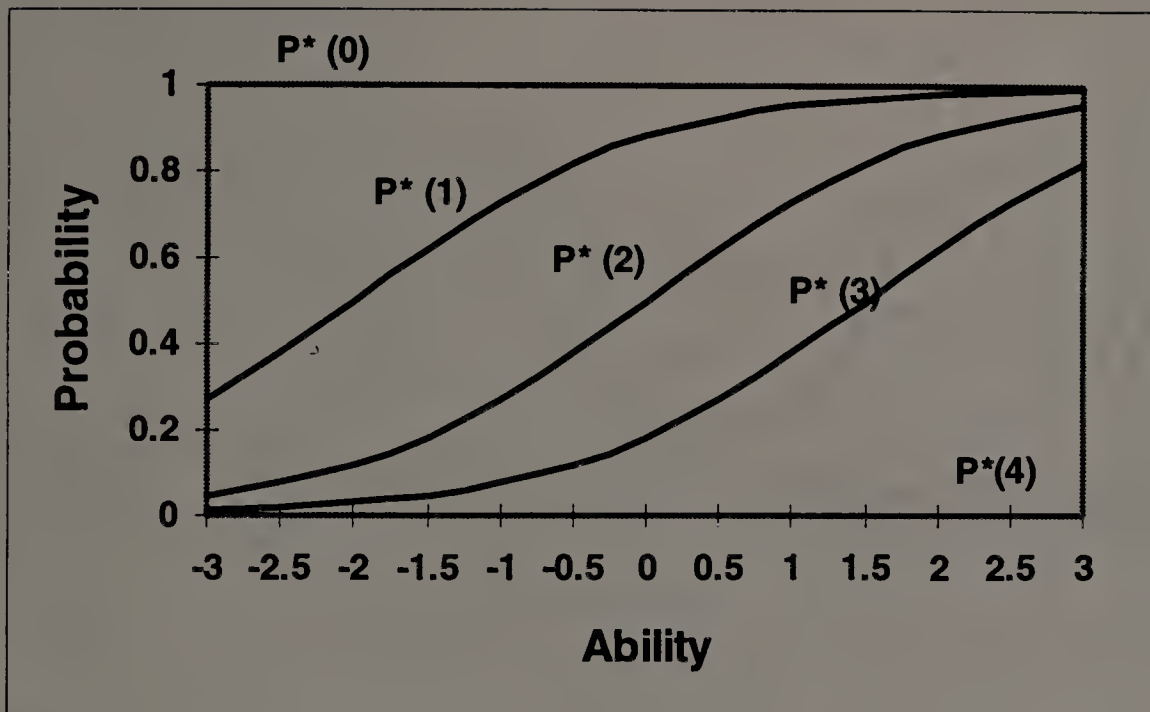


Figure 1. Boundary characteristic curves for four category item

($a=1$, $b_1=-2$, $b_2=0$, $b_3=1.5$)

In Figure 1, P^*_{j1} specifies the probability of responding in category 1,2, or 3 rather than category 0, P^*_{j2} specifies the probability of responding in category 2 or 3 rather than category 0 or 1, and P^*_{j3} specifies the probability of responding in category 3 rather than category 0,1, or 2.

The second stage in obtaining the probability that an examinee will respond in a given category k is subtracting adjacent category characteristic curves. Samejima (1969) defined the probability that an examinee would respond in a given category as

$$\begin{aligned}
 P_{jk}(\theta) &= P^*_{jk}(\theta) - P^*_{jk+1}(\theta) \\
 &= \frac{1}{1+\exp[-a_i(\theta-b_{ik})]} - \frac{1}{1+\exp[-a_i(\theta-b_{ik+1})]}
 \end{aligned}
 \tag{4}$$

where, $k=0,1,\dots,m$ and the probability of responding in category 0 or above, $P_{j0}^*(\theta)=1.0$ and the probability of responding in the highest category $m+1$, $P_{j(m+1)}^*(\theta)=0$.

For example, the probabilities of responding in each of the four categories 0, 1, 2, and 3 are obtained by employing the operating characteristic curves (P_{jk}):

$$P_{j0}(\theta) = P_{j0}^*(\theta) - P_{j1}^*(\theta) = 1.0 - P_{j1}^*(\theta),$$

$$P_{j1}(\theta) = P_{j1}^*(\theta) - P_{j2}^*(\theta),$$

$$P_{j2}(\theta) = P_{j2}^*(\theta) - P_{j3}^*(\theta), \text{ and}$$

$$P_{j3}(\theta) = P_{j3}^*(\theta) - P_{j4}^*(\theta) = P_{j3}^*(\theta) - 0.0 = P_{j3}^*(\theta).$$

The operating characteristic curves given by equation 4 for this example of the GRM are presented in Figure 2. As can be seen, the probability of responding in either of extreme categories is a monotonic function, while the probability of responding in any of the other categories is a nonmonotonic symmetric function.

In general, a boundary curve P_{jk}^* can be reduce the graded scored item to a dichotomously scored item. That is, the graded responses can be classified into two categories; scores lower than k and scores equal to or greater than k , for $k=0,1,2,\dots, m-1$. The equation 4 can be the dichotomous two-parameter logistic IRT model, if an item has two response categories.

$$\begin{aligned} P_{j0}(\theta) &= P_{j0}^*(\theta) - P_{j1}^*(\theta) = 1 - P_{j1}^*(\theta) \\ P_{j1}(\theta) &= P_{j1}^*(\theta) - 0.0 = P_{j1}^*(\theta) \end{aligned} \tag{5}$$

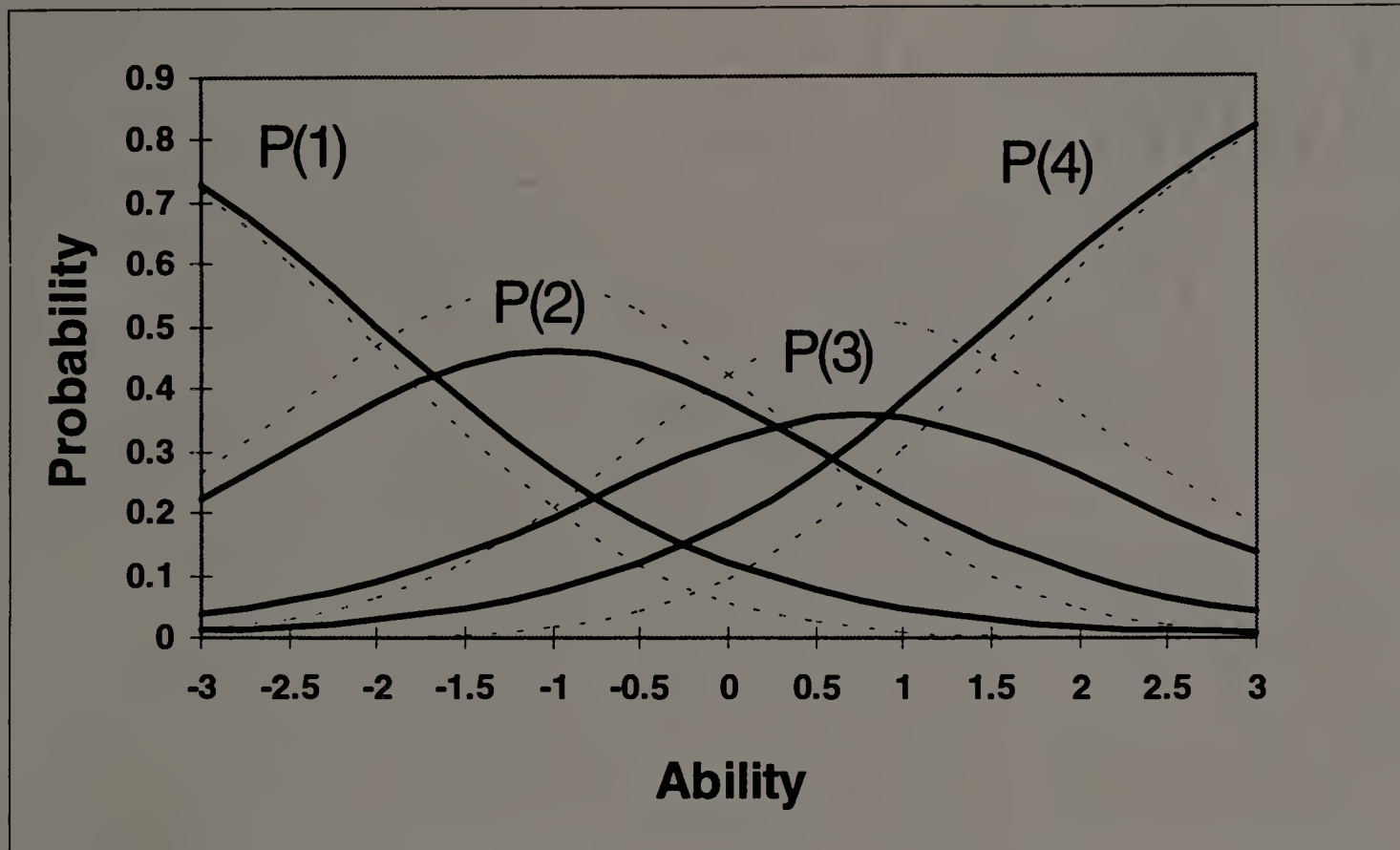


Figure 2. ICCCs with four ordinal responses under the GRM and the PCM ($a=1$, $b_1=-2$, $b_2=0$, $b_3=1.5$, ___ : the GRM and : the PCM)

Substituting the equality for P_{ji}^* from equation 3, P_{ji} can be rewritten

$$P_{ji}(\theta) = \frac{e^{a_j(\theta-b_j)}}{1 + e^{a_j(\theta-b_j)}} \quad (6)$$

The difference between boundary characteristic curves and operating characteristic curves for the GRM are apparent, when it is plotted on the same graph, as in figures 1 and 2. Figure 1 represents boundaries on the cumulative probabilities of response categories for a four-category item. While there are four response categories in the example, there are only three boundaries between categories as well as an upper boundary of one and a lower boundary of zero.

Figure 2 represents item response characteristic curves or operating characteristic curves. It depicts the probability of responding to category k at the ability level θ . The person with low ability has a high probability at the "lowest" category (score=0), the person with middle ability has moderate probability at any of four categories, and the person with a high ability has a high probability at the highest category.

The boundary characteristic curves representing cumulative probabilities can be characterized by the parameters a_j and b_{ij} . Since these curves represent the sum of the response category probabilities, negative differences between curves are not possible. Thus, these boundary curves cannot cross. The boundary characteristic curves are assumed to have the same discrimination parameter a_j in the GRM, but there is no requirement that the discrimination parameter is the same in all items. The discrimination parameter a_j poses no interpretive difficulties since it is the same for all item response categories. The value of a_j has the same meaning as in dichotomous IRT models.

However there is a problem in interpreting the boundary parameters of the operating characteristic curves because there is one less boundary parameter than item response categories due to the restriction of $\sum_{k=0}^m P_k(\theta) = 1$. In the GRM, boundary parameter, b_{jk} , is defined as the ability level which corresponds to the point of inflection of the category characteristic curve, the P_{jk}^* . Samejima(1969) showed that the modal point of an operating characteristic curve was given by $b'_{jk+1} = (b_{jk} + b_{jk+1})/2$ except for the first and the last response categories. So, the first and last boundary parameters of an item response category retains their interpretation as the point on the ability scale at

which the probability that the response will be allocated to their category is .5: $P_{j0}(\theta) = P_{jm}(\theta) = 0.5$. The parameter b'_{jk} can be interpreted in a manner analogous to the difficulty parameter of dichotomously scored items. The boundary parameter in the category curves must be ordered $b_m > b_{m-1} > \dots > b_1$, so the location parameters in the operating characteristic curves may also be ordered $b_m > b'_{m-1} > b_{m-1} > \dots > b_1$. There is no requirement that location parameters be equally spaced, only that they be monotonically decreasing or increasing.

2.4.1.2 The Partial Credit Model

The partial credit model (PCM) presented by Masters (1982) also assumes that responses to an item are ordered like the GRM, however there are differences between the PCM and the GRM on formulation of the operating characteristic function and interpretation of the parameters, since the two models have different structures of parameter formation: the PCM is an extension of Rasch's (1960) dichotomous model and the GRM is an extension of two parameter item response model.

Masters (1982) classified ordered level of responses into four types: (1) Repeated trials data results when respondents are given a fixed number of independent attempt at each item on a test. The observation x is the number of successes on the item and takes values from 0 to m . This format is useful for tests of psychomotor skills in which the observation is a count of the number of items in m attempts that a task is successfully performed; (2) Count data results when there is no upper limit on the number of independent successes (or failures) a person can make on an item. Under this format

observation x may be a count of the number of times a person completes a task in a specified period of time, or a count of the errors a person makes in reading a passage on an oral reading test; (3) Rating scale data is a fixed set of ordered response alternatives used with every item. The format of response alternatives can be used for Likert-type statements; (4) Partial credit data comes from an observation format which requires the prior identification of several ordered levels of performance on each item and thereby awards partial credit for partial success on items. The motive for partial credit scoring is the hope that it will lead to more precise estimate of a person's ability than a simple pass/fail score.

Masters (1982) developed the partial credit model (PCM) for the analysis of partial credit data. Partial credit types of data needs several levels of performance to complete an item. For example, an item involves four levels of performance, where 0 denotes no response and 3 a successful completion. Masters (1982) presented an example as follows:

$$(7.5/0.3 - 16)^2 = ?$$

Step 1: evaluate the quotient $7.5/0.3$

Step 2: Subtract 16 from the result of step 1.

Step 3: Square the result of step 2.

To complete this item, certain ordered steps must be performed correctly. Under the PCM the necessary order is not relative difficulties of steps but the steps that must be taken to completed. That is, it is impossible to succeed at the second step without completion of the first step. Masters (1982) interpreted the ordered category scores for an

item to represent the number of subtasks or steps in the item that has been successfully completed.

Masters (1982) extended the logistic Rasch model to the PCM and is shown as

$$\begin{aligned}
 P_{jl}^*(\theta_i) &= \frac{P_{jl}(\theta_i)}{P_{j0}(\theta_i) + P_{jl}(\theta_i)} \\
 &= \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}
 \end{aligned}
 \tag{7}$$

This specifies the probability of person i succeeding on item j given that only two outcomes are possible. The item difficulty b_j is rewritten b_{j1} to make it explicit that this is the difficulty level associated with completing the first step in item j . Therefore, in a four category item, 3 equations for each step are used; P_{1ni}^* , P_{2ni}^* , P_{3ni}^* in the PCM, while in the GRM there are four boundary characteristic curves; P_{0ni}^* , P_{1ni}^* , P_{2ni}^* , P_{3ni}^* .

Under the PCM, completing the k -th step means choosing the k -th response alternative over the $k-1$ th response alternative. That is, the probability of getting a score k on the item rather than $k-1$ is given by,

$$P_{jk}^*(\theta_i) = \frac{P_{jk}(\theta_i)}{P_{jk-1}(\theta_i) + P_{jk}(\theta_i)} = \frac{\exp(\theta_i - b_{jk})}{1 + \exp(\theta_i - b_{jk})}
 \tag{8}$$

Since an examinee must earn one of all possible scores, the following equation (9) holds:

$$P_{j0}(\theta_i) + P_{j1}(\theta_i) + \dots + P_{jk}(\theta_i) = 1
 \tag{9}$$

By using (8) and (9), Masters (1982) arrived at the general expression (10) for the probability of examinee i getting a score k on item j ,

$$P_{jk}(\theta_i) = \frac{\exp \sum_{s=0}^k (\theta_i - b_{js})}{\sum_{v=0}^{m_j} \exp \sum_{s=0}^v (\theta_i - b_{js})}, \quad k=0,1,\dots,k,\dots, m_j \quad (10)$$

where $b_{j0}=0$ and $\sum_{k=0}^0 (\theta_i - b_{jk})=0$. The numerator contains only terms for the step completed, and the denominator is the sum of all possible numerator terms.

2.4.1.3 Comparison of the GRM and the PCM

Both GRM and PCMs are appropriate for items in which responses to an item can be classified into m ordered categories, but there are differences between the GRM and the PCM in terms of interpretation of parameters. The differences in interpretation is due to the assumptions underlying the derivation of the operating characteristic curves and the employment of different characteristic functions for an item to obtain the probability that an individual will respond in a given category k . The PCM models the probability of person scoring x on the category k in an item j as a function of the person's position of the trait on the variable and the difficulties of the steps in the item j . Therefore in the PCM a step difficulty (b_{jk}) is associated with only the step. In contrast, in the GRM, a boundary difficulty (b_{jk}) is related to the other categories because the GRM is structured according to cumulative probabilities (the probability of person scoring x in or above the category k

in an item j). Consequently, the probability in any category is the difference between successive cumulative probabilities.

In the GRM, the category boundary parameter b_{jk} associated with a given category score k is defined as the ability level which corresponds to the point of inflection of the boundary characteristic curve P^*_{jk} (not the item curve characteristic curve P_{jk}), i.e., the category boundary parameter b_{jk} is the ability level where the probability of responding in categories greater than or equal to category score k ($P^*_{jk}(\theta_i)=0.5$). The definition of category boundary parameters requires that the category boundaries be ordered ($b_k > b_{k-1}$).

In contrast, in the PCM, the step parameter b_{jk} is defined as the ability level where the probability of responding in category k equals the probability of responding in category $k-1$, i.e., the step parameter b_{jk} is the point of intersection of adjacent category characteristic curves ($P_{j,k-1}(\theta_i)=P_{jk}(\theta_i)$). For example, b_1 is the point of intersection of $P_1(\theta_i)=P_2(\theta_i)$ in figure 2. Since the probability of responding in categories other than k and $k-1$ are not taken into consideration in the definition of the step difficulty, the difficulties of the previous step or later steps have no bearing on the difficulty of the step associated with the category score k . Thus, the PCM requires that the steps be ordered, but the step difficulties not to be ordered.

In the PCM, a category score indicates the number of successfully completed steps. The more steps successfully completed the larger a category score; a higher category score indicates greater ability than does as a lower category score. In the GRM there are boundary scores (instead of a step) above which a person is expected to obtain a

certain category score. If there are four potential category scores (0,1,2,3), the probabilities correspond to the probability of obtaining scores of 1, 2, or 3 over a score of 0, the probability of obtaining scores of 2 or 3 over scores 0 or 1, or the probability of obtaining a score of 3 over scores 2, 1, or 0. A higher category indicates greater ability in the GRM and PCM. It is important to note, however, that in the GRM, b_{jk} are always ordered such that $b_k > b_{k-1}$, while in the PCM steps need to be ordered but not necessarily by their difficulties.

The GRM and the PCM differ additionally in their treatment of the item discrimination index. The PCM assumes items in a test (or inventory) all have equal discrimination powers, while the GRM allows items in a test to differ in terms of their ability to discriminate among examinees of different levels. As a result of this, in the PCM a raw score is the sufficient statistics for the ability parameter. Hence, everyone who has the same total number of steps completed successfully on the test will receive the same ability estimate as in the dichotomous Rasch model, even though the specific steps completed on individual items may differ and the steps may be of widely varying difficulty.

2.4.1.4 The Rating Scale Model

The rating scale model (RSM) was designed by Andrich (1978) for instruments in which Likert-type statements were used to measure attitude. He presented the model using the Rasch model in the context of analysis of rating scales having an ordinal response category scale. In a rating scale, however, ordinal response levels are not

defined by a series of item subtasks, but by a fixed set of ordered points within items. As the same set of rating points is used with every item, the relative difficulties of the steps within each item does not have to vary greatly from item to item. Therefore the category coefficients and the scoring functions in Rasch's general model are interpreted in terms of thresholds on the latent continuum and discriminations at the thresholds. The RSM is a special case of PCM.

Masters (1982) decomposed the item step difficulty of the PCM into two components: $b_{jk} = b_j + \tau_k$, where b_j is location (or scale value) of item j and τ_k is the location of the k -th category in each item relative to that item's scale value. The τ_k s are also known as thresholds because they separate the $m+1$ ordered categories. By substituting the above equation of the item step difficulty, $b_{jk} = b_j + \tau_k$, into the PCM (equation 8), the RSM (Andrich, 1978) is obtained as equation (11):

$$P_{jk}(\theta_i) = \frac{\exp \sum_{s=0}^k [\theta_i - (b_j + \tau_s)]}{\sum_{k=0}^m \exp \sum_{s=0}^k [\theta_i - (b_j + \tau_s)]} \quad (11)$$

Equation (11) can be reexpressed as

$$P_{jk}(\theta_i) = \frac{\exp[-\sum_{s=0}^k \tau_s + k(\theta_i - b_j)]}{\sum_{k=0}^m \exp[-\sum_{s=0}^k \tau_s + k(\theta_i - b_j)]} \quad (12)$$

$$= \frac{\exp[K_k + k(\theta_i - b_j)]}{\sum_{k=0}^m \exp[K_k + k(\theta_i - b_j)]}$$

where k indicates the number of thresholds passed, $\theta(i)$ is the person's latent trait (i.e., attitude), b_j is the item scale value, and k_k is defined as

$$K_k = -\sum_{s=0}^k \tau_s$$

for $k=1,2,\dots,m$; $k_k=0$ when $k=0$.

As a result of simplification of the equation $b_{jk} = b_j + \tau_k$, the τ_k s are constant across items and need to be estimated once for the entire item set, however the item scale values b_j are estimated individually for each item (Andrich, 1978; Masters, 1982). When this model is applied to the analysis of a rating scale, a position on the variable $\theta(i)$ is estimated for each person, a scale value b_j is estimated for each item j , and m response thresholds $\tau_1, \tau_2, \dots, \tau_m$ are estimated for the $m+1$ rating categories. As was the case with the PCM, the Andrich's RSM assumes items are equally effective at discriminating among examinees.

2.4.1.5 The Generalized Partial Credit Model

The generalized partial credit model (GPCM) extended by Muraki (1992) is emerging as a popular model for ordinal response items because it has advantages of both the GRM and the PCM. The GPCM, like the PCM, is an extension of the Rasch formulation to polytomously scored items and, like the GRM, allows the discrimination index for each item and category difficulty indices for categories in each item.

In the GPCM, the probability of a person with trait level θ responding in category k ($k=1,2,\dots,m_j$) on item j is defined as

$$P_{jk}(\theta) = \frac{\exp[\sum_{v=1}^k a_j(\theta - b_{jv})]}{\sum_{c=1}^{m_j} \exp[\sum_{v=1}^c a_j(\theta - b_{jv})]}, \quad k=1, 2, \dots, m_j \quad (14)$$

Here, a_j is the discrimination or slope parameter for item j and b_{jk} , the item category threshold parameter, is the step difficulty for the k -th step of item j . In the GPCM a_j is interpreted as the degree to which categorical response varies among items as ability level changes. If the slope parameter a_j is changed from 1.0 to 0.5, the intersection points b_{jk} (step difficulties) of all ICCCs are unchanged but the curves become flatter. Note that b_{j1} is arbitrary and may be defined as 0.

Following Andrich (1978) and Muraki (1992) decomposed the category parameter b_{jk} into two components, b_j and d_k such that $b_{jk} = b_j - d_k$. This decomposition is appropriate for rating scales. The parameter b_j is the item location parameter, and d_k is the

category threshold parameter. In the following, b_{jk} will be designated as the *step difficulty* parameter, b_j , *the item location* parameter, and d_k , the category threshold parameter.

The GPCM is formulated using the same assumption as the PCM that the probability of choosing category k over a category $(k - 1)$ in an ordinal response item is governed by the dichotomous response model. Completing step k means choosing response alternative k over response alternative $(k - 1)$. In the GPCM, an examinee's choice among successive categories (k) is represented as a series of steps, completed in order, but the step difficulty parameters (b_{jk}) of the successive categories need not be ordered. Since the step parameter, b_{jk} , is defined as the ability level where the probability of responding in category k equals the probability of responding in category $k-1$, the values of step parameters represent the relative magnitude of the adjacent probabilities of P_{jk} and P_{jk-1} .

2.4.2 Model for Nominal Responses

2.4.2.1 The Nominal Response Model

For a test item in which response options are not necessarily ordered, a nominal response model is appropriate. Bock (1972) employed the multivariate logistic function which was a generalization of the bivariate logistic function derived by Gumbel (1961) to get operating characteristics for each response category of a nominally scored item.

The nominal response model (NRM) provides a direct expression for obtaining the probability of an examinee with ability θ responding in the k -th category of item j .

The mathematical form of the multivariate logistic function is equation (15) or the NRM is

$$P_{jk}(\theta) = \frac{e^{Z_{jk}(\theta)}}{\sum_{k=1}^{m_j} e^{Z_{jk}(\theta)}}, \quad (14)$$

where $Z_{jk}(\theta) = a_{jk}\theta + c_{jk}$, $k=1, 2, \dots, m_j$.

Because of the indeterminacy in the model, it is necessary to impose a linear constraint on the item parameters, $\sum_k Z_{jk}(\theta) = 0$ where $k=1, 2, \dots, m_j$.

Unlike ordinal response models, under the NRM an examinee's total score can not be summed and the response score received by the examinee for an item has no meaning other than to designate the response category. That is, item response category characteristic curves (IRCCCs) for the NRM, $P_{jk}(\theta_i)$, just depicts proportion of responses assigned to each of the nominally scored response categories as a function of ability.

In the NRM, a_{jk} is considered the slope (discrimination) parameter and c_{jk} is the intercept parameter of the nonlinear response function associated with the k -th category of item j , while m_j is the number of categories of item j (i.e., $k=1, 2, \dots, m_j$). In the NR model each category's ability to discriminate among examinees is captured by the category's individual discrimination parameter, a_{jk} . The a_{jk} is analogous to and has an interpretation similar to a traditional discrimination index. That is, a category with a large a_{jk} reflects a response pattern where as one progresses from the lower ability groups to the higher ability groups there is a corresponding increase in the number of persons

who answered the item in that category, and for categories with negative a_{jk} s this pattern is reversed.

Generally, large values of c_{jk} are associated with the categories with large frequencies. As the value of c_{jk} becomes increasingly small, the frequencies for the corresponding categories decrease.

CHAPTER 3

REVIEW OF THE LITERATURE

3.1 Estimation Procedures for IRT Models

Lord (1952) and Birnbaum (1968) developed joint maximum likelihood estimation in item response models. Item and ability parameters are unknown in a typical testing situation, and hence both item and ability parameters have to be estimated simultaneously. Since both item and ability parameters are unobservable, in order to obtain the estimates of item (or ability) parameters, the number of examinees (or items) must be increased. The joint MLE of item and ability parameters are not consistent when both sets of parameters have to be estimated simultaneously (Swaminathan, 1982).

This problem can be solved by integrating with respect to the incidental parameters (ability parameters) if they are assumed to be continuous or by summing over their values if they are discrete. The resulting likelihood function is the marginal maximum likelihood function. The marginal maximum likelihood (MML) estimators of the parameters are those values that maximize the marginal likelihood function. MML estimators possess several useful and important properties. Under usual circumstances, the MML estimators are consistent, i.e., asymptotically unbiased; efficient, i.e., asymptotically the estimators have the smallest variance; and asymptotically normally distributed (Swaminathan, 1983).

MMLE procedure was applied to estimate parameters in IRT models by Bock and Lieberman (1970). They provided MML estimators of the two-parameter item response

model and assumed that the ability distribution was normal with zero mean and unit variance and integrated over θ numerically. They obtained stable parameter estimates for few items using the Gauss-Hermite quadrature procedure to perform the necessary integration. However their method had a computational problem in the case of the large number of items, because the likelihood function had to be evaluated for all possible response patterns. This restricts the practical application of the procedure to approximately 10 item tests (Bock & Lieberman, 1970).

Bock and Aitkin (1981) solved the computational difficulties of the Bock and Lieberman procedure by characterizing the distribution of ability empirically and employing a modification of EM algorithm formulated by Dempster, Laird, and Rubin (1977). The MMLE with EM algorithm has been implemented in the computer program BILOG (Mislevy and Bock, 1986), MULTILOG (Thissen, 1991), and PARSCALE (Muraki, 1993) to obtain parameter estimates of dichotomous and polytomous IRT models.

3.2 MMLE Procedure for Polytomous IRT Models

For the polytomous response models, let U_{jki} represent an element in the matrix of the observed response pattern i . $U_{jki}=1$ if the response to item j is in the k -th category, otherwise $U_{jki}=0$. The probability of an examinee in the pattern i obtaining the response vector, U_{jk} , is

$$P(U_{jk} | \theta_i) = \prod_{j=1}^n \prod_{k=1}^m P_{jk}^{U_{jk}} \quad (15)$$

where, $P_{jk}(\theta)$ is probability that an examinee i responds category k in item j . The marginal probability of the observed response pattern i is

$$P_i(U_{jk}) = \int_{-\infty}^{\infty} P(U_{jk} | \theta_i) g(\theta) d\theta \quad (16)$$

where $g(\theta)$ is the population distribution of ability for examinees. There are m^n response patterns in for n items with m categories. If r_i denotes the number of examinees obtaining response pattern i and N is the total number of examinees sampled from population, the likelihood function is given by

$$L = \frac{N!}{\prod_{i=1}^{m^n} r_i!} \prod_{i=1}^{m^n} P_i(U_{jk})^{r_i} \quad (17)$$

Taking the natural logarithm of likelihood function yields

$$\ln L = [\ln N! - \sum_{i=1}^{m^n} \ln r_i!] + \sum_{i=1}^{m^n} r_i \ln P_i(U_{jk}) \quad (18)$$

The MML estimators are obtained by differentiating $\ln L$ with respect to each parameter, setting the derivatives equal to zero, and solving the equations.

3.3 Previous Research on MMLE Procedure

Several studies (Bock & Aitkin, 1981; Drasgow, 1989; Mislevy & Stocking, 1989; Seong, 1990; Thissen, 1982; Stone, 1992; Yen, 1987) have investigated the accuracy of MMLE parameters for the dichotomous IRT models. Thissen (1982) has adapted MMLE with EM algorithm to the Rasch model and showed that the results was comparable to that of conditional estimation procedure. Yen (1987) and Mislevy & Stocking (1989) compared the computer program BILOG with LOGIST and provided some guidelines for using these programs.

Drasgow (1989) evaluated MML estimates for the two parameter logistic model using parameter values of Job Descriptive Index (JDI; Smith, Kendall, & Holin, 1969). The results showed that MMLEs were far more accurate than JMLEs. While for items with less extreme values of parameters, as few as 200 examinees and 5 items were required for providing unbiased parameter estimates with reasonably small SEs, 500 examinees and 10 items were required for items with extreme values of parameters ($a < 0.8$, $a > 1.40$; $|b| > 1.50$). He pointed out the accuracy of estimation depended on the values of the item parameters and suggested using appropriate Bayesian prior distribution for extreme values of the parameter.

Seong (1990) studied the effect of ability distributions on robustness of the MML estimates for the two-parameter logistic model. Appropriate specification of the ability distribution increased the accuracy of estimation for item and consequently the ability parameters when the sample size was large. With a small sample size (100 examinees), the result for item parameter estimation was inconsistent with that of a large sample size

(1000 examinees). That is, the item parameter estimates in cases of the matched distribution were less accurate than those of the non matched ability distributions.

Stone (1992), in extending to Seong's (1990) study and Drasgow's (1989) study, examined the effect of test lengths and the ability distribution on the item parameter estimation. Even with the small sample size (250) and a short test (10 items) item difficulty estimates were stable and precise regardless of ability distribution, but item discriminate estimates were stable and precise only when the true distribution of ability was normal.

Test length had a major effect on discriminate parameter estimates. As test length increased from 10 to 40 items, bias in estimates of discrimination parameters was reduced even under nonnormal distribution of ability. Root mean square error (RMSE) for estimates of discrimination parameters was rapidly reduced when test items increased from 10 to 20 regardless of ability distributions.

Stone (1990) also found the value of item parameters affected the accuracy of parameter estimation. Stone (1990) selected three discrimination parameters; low ($a=0.8$), medium ($a=1.9$), and high ($a=3.0$), and three difficulty parameters; average ($b=-0.02$), easy ($b=-2.68$), and hard ($b=1.8$). For the low discriminating item, bias was negligible irrespective of the number of test items, the true ability distributions, and sample size. For the average and high discriminating items bias was greater when the distribution of ability deviated from $N(0,1)$ for the test comprised of 10 or 20 items, irrespective of sample size. For the average difficulty item bias was negligible irrespective of the number of test items, the true ability distributions, and sample size. For

the easy and hard items bias was higher with non-normal distribution regardless of test lengths.

In addition, different combination of a and b parameter values affected the parameter estimates. For example, the smallest RMSE was observed for the average difficulty parameter ($b=-0.02$) and low discriminate parameter ($a=0.8$). Greater RMSE were observed for the highly discriminating item and the extremely easy item.

Compared to the research on dichotomous IRT models, there is very little research dealing with the parameter estimation under polytomous IRT models. Reise and Yu (1990) have studied the effects of sample sizes, ability distributions, and the range of discrimination parameters on the accuracy of parameter estimates for the GRM using computer program MULTILOG (Thissen, 1986). They studied sample sizes of 250, 500, 1000, and 2000, normal, uniform, and negatively skewed distribution of examinees' ability, and high, middle, and low discrimination parameters with a fixed five category-25 item test. They found that all three factors included in the study affected the accuracy of item parameter estimates.

Reise and Yu (1990) reported that uniform ability distribution conditions were slightly superior on average accuracy of item parameter estimates compared to normal and skewed ability distributions. However the RMSE and correlation results for all separate 36 conditions displayed unreasonably large RMSE and low correlations under normal and skewed ability distributions with small sample size (250 examinees). Especially RMSE and correlations of item category parameter estimates (b_1, b_2, b_3 , and b_4) are much larger than those of item discrimination parameter estimates. They also

found that RMSE for extreme value of category parameters (b1 and b4) was larger than the middle categories' (b2 and b3); correlation between true values and estimates showed similar patterns. This result can be attributed to the sample size in each category. Since polytomous IRT models have more categories than dichotomous IRT models, and more categories may have extreme values of category parameters, each category is more affected by the number of examinees at each ability level and the locations of the item categories relative to the ability distribution than the dichotomous case. In addition, they showed that true a value affected the average RMSE for difficulty parameter estimates (b) and the average correlation for discrimination parameter estimates (i.e., with high a value RMSE for b was small and correlation for a was high).

Walker-Barnick (1990) investigated the accuracy of the parameter estimates for the PCM of Masters (1982). Factors in the study were the ratio of sample size to the number of parameters to be estimated (1:1, 2:1, 3:1, and 4:1), the number of categories in items (4 and 5), and distribution of the examinees' ability. The computer program MSTEPS (Wright et al., 1988) with a joint maximum likelihood estimation procedure was used in the study. Walker-Barnick (1990) showed that the parameter estimates of the PCM were stable under all conditions. The results of the study cannot be generalized because the study used a long test (80 items) with moderate difficult items.

De Ayala (1995) examined the effect of the ratio of the sample size to item parameters to be estimated, distribution of examinees' ability, the amount of information of item, and the number of categories in items on the NRM by Bock (1972) using computer program MULTILOG (Thiseen, 1991) with a fixed 28 test length. It was found

that ability distribution, sample size, and item information affected the accuracy of the parameter estimates. The results showed that as the latent trait distribution departed from a uniform distribution, the accuracy of estimating the discrimination parameter decreased. This result consistent with that of Reise & Yu (1990).

De Ayala (1995) pointed out that the effects of the form of ability distribution on RMSE, in part, may be attributed to the distribution of responses across item categories. It was found that the uniform distribution produced the greatest dispersal of responses across item categories and that the positively skewed distribution produced least variability in the examinees' responses. Therefore, if there are insufficient number of examinees responding to a particular item category, then that category will not be as accurately estimated as other categories that have a large number of responses.

Choi, Cook, and Dodd (1996) investigated the effect of the sample size, the number of categories in each item, and the test length on the recovery of parameters for the PCM using MULTILOG (Thissen, 1991) computer program. They found that sample size and the number of categories were the most important factors that affected the accuracy of item parameter estimates. They pointed out as the number of categories increased, the sparsity of the observation in the extreme category was magnified and affected estimation. They further showed that given a fixed sample size, adding more items slightly decreased the accuracy of estimation for 7 category items, while it increased the accuracy of estimates for the 4 category items. They concluded that test length did not significantly impact on the accuracy of MMLE item parameter estimation. However they did not use the same number of items for the 4 and 7 category tests because

they used the ratio of sample size to parameters to be estimated as a variable of study.

Since the number of parameters is a function of the test length and the number of categories, they could not examine the effect of test lengths and the number of categories simultaneously.

3.4 Bayesian Estimation of Parameters in Polytomous IRT Models

A problem that is often encountered with the MMLE is that the item parameter estimates drift out of bounds. One way to resolve this problem is to restrict particular values for the parameters. However rather than imposing arbitrary restrictions on the parameter estimates, a Bayesian can be employed by incorporating prior knowledge about the parameters.

The prior probability and likelihood function can be combined using Bayes' theorem. The resulting posterior distribution contains all the information about the parameters of interest. Bayesian approaches in IRT can be distinguished by whether item parameter estimation takes place with or without marginalization over ability parameters. If marginalization is not used, the approach is called joint Bayesian estimation; if marginalization is used, the approach is called marginal Bayesian estimation.

Swaminathan and Gifford (1982, 1985, and 1986) employed a joint Bayesian estimation procedure to estimate parameters of the dichotomous item response models. They implemented the hierarchical Bayes procedures for the specification of prior beliefs following the approach taken by Lindley (1971) and Lindley and Smith (1972). They found that different specifications of prior distributions had relatively modest effects on

the Bayesian estimates except using extreme prior, and using any prior improved the accuracy of estimates. The accuracy of estimation in b and ability parameters did not seem to be affected by the specification of prior information, whereas a and c parameters were affected by the specification of prior. For the a parameter, the Bayesian procedure produced smaller error than JMLE because the priors arrested the outward drift of the estimates (Gifford & Swaminathan, 1990).

A marginalized Bayesian procedure was implemented in computer program BILOG by Mislevy and Bock (1986) for estimating item parameters. Mislevy and Bock imposed the lognormal prior distribution on the discrimination parameters as the default in BILOG. A normal prior distribution may be specified for the location parameters but using the prior is optional in BILOG. Evidence presented by Swaminathan and Gifford (1985) indicates that specification of non-informative priors for the location and ability parameters with an informative prior for the discrimination parameter appears to be reasonable approach because when an informative prior is specified for the discrimination parameter, the estimation of all the parameters proceeds smoothly.

Several studies have investigated the performance of prior distributions with MMLE for the dichotomous item response models (Harwell & Janosky, 1991; Lim & Drasgow, 1990; Mislevy, 1986; Yen, 1987). Results of all of these studies showed that using prior distributions for parameters provided more accurate estimates than without using prior distribution. Harwell and Janosky (1991) examined the effect of small number of examinees and items, and different variances for the prior distributions of discrimination parameters on item parameter estimation in BILOG using two-parameter

model. They found that for test of 15 and 25 items, the effect of the variance of prior distribution was negligible when 250 or more examinees were available. For smaller samples (i.e., 75, 100, and 150) and a short test (i.e., 15 items), the variance of prior distribution played a prominent role in the quality of item parameter estimation.

Similar procedures of imposing prior distributions are employed for the polytomous IRT models. For the polytomous IRT models, let U_{ijk} represent an element in the matrix U of the observed response pattern for examinee i . $U_{jki} = 1$ if the response of examinee i to item j is in category k , otherwise $U_{jki} = 0$. Further assume that the latent space is unidimensional and that the conditional probability of a response pattern i , for m response categories and n items, given θ and item parameters a_{jk} and b_{jk} , is the joint probability:

$$P(U_i | \theta_i, a_{jk}, b_{jk}) = \prod_{j=1}^n \prod_{k=1}^m P_{ijk}^{U_{ijk}} \quad (20)$$

where P_{ijk} is the probability that examinee i responds in category k on item j . The marginal probability of the observed response pattern i is

$$P(U_{ijk} | a_{jk}, b_{jk}) = \int_{-\infty}^{\infty} P(U_{ijk} | \theta_i, a_{jk}, b_{jk}) g(\theta) d\theta, \quad (21)$$

where $g(\theta)$ is the population distribution of ability for examinees. The marginal probability of obtaining the response pattern matrix U is then given by

$$P(U | a_{jk}, b_{jk}) = \prod_{i=1}^N P(U_i | a_{jk}, b_{jk}) \quad (22)$$

Once the observations are made, this becomes the likelihood function of the parameters, given by

$$L(U | a_{jk}, b_{jk}) = \prod_{i=1}^N \prod_{j=1}^n \prod_{k=1}^m P_{ijk} . \quad (23)$$

According to Bayes's theorem the posterior probability distribution for item parameters given the data is proportional to the product of the likelihood function and the prior distribution of the item parameters, i.e.,

$$P(a_{jk}, b_{jk} | U) \propto L(U | a_{jk}, b_{jk}) P(a_{jk}, b_{jk}) . \quad (24)$$

The joint probability $P(a_{jk}, b_{jk})$ is the joint prior distribution of the vectors of item parameters and is an expression of the prior belief or information the investigator has been regarding these parameters. In the first stage of the model, we assume *a priori* that the parameters a_{jk} and b_{jk} are independently distributed, i.e., $P(a_{jk}, b_{jk}) = P(a_{jk}) P(b_{jk})$. The next step in Bayesian inference is to specify a prior distribution for each item parameter.

The computer program PARSCALE that is designed to obtain estimates of parameters in the GPCM, transforms the slope parameters into new parameters, $\alpha_j = \log a_j$. It assumes that each a_j has a lognormal prior distribution over $0 \leq a_j \leq \infty$. This implies that $\alpha_j = \log a_j$ has a normal prior distribution with a density that is proportional to $\exp\{-1/2[(\alpha_j - \mu_\alpha) / \sigma_\alpha]^2\}$. The normal prior distribution of each α_j is defined by its parameters, α_j and μ_α , which are assigned default values of 0 and 0.5, respectively, by

the PARSCALE program. The default values of $\mu_\alpha = 0$ and $\sigma_\alpha = 0.5$ in PARSCALE results in $\mu_a = 1.13$ and $\sigma_a = 0.6$.

It is not possible to specify a prior distribution for the step difficulty parameter b_{jk} in PARSCALE. Instead, a normal prior with mean μ_b and standard deviation σ_b is specified for the threshold parameters (b_j) with default specifications of $\mu_b = 0$ and $\sigma_b = 2$. Thus, default as well as user-provided priors can be specified for the slope and the threshold parameters.

3.5 Summary

In this chapter, MML and Bayesian procedures for IRT models and the previous research which examined factors influencing parameter estimation of dichotomous and polytomous IRT models were described. The research on MML estimation with dichotomous IRT models has indicated that sample size, ability distributions, test lengths, the value of item parameters, and the combination of discrimination and difficulty parameter values affected parameter estimation. These factors also affect estimation in polytomous IRT models while with dichotomous IRT models, ability distribution had an effect in the estimation of c parameter (Swaminathan & Gifford, 1986), the ability distribution plays an important role for the estimation of item parameters in polytomous IRT models. Baker (1987) indicated that with a non uniform ability distribution an interaction occurred between the number of examinees at each ability level and the estimated parameters of the ICC. For reason of this is polytomous IRT models have more categories than dichotomous IRT models and hence require more observations in each

category for the accurate estimation of parameter. Despite the availability of some research on parameter estimation in polytomous IRT models, considerable research needs to be completed especially with respect to the interaction among the factors mentioned above and their effect on estimation.

In addition, the research on Bayesian procedure with dichotomous IRT models showed that Bayesian procedure was superior to the maximum likelihood procedures in that estimates remained in the parameter space, were more accurate, at least in small samples and less biased (Gifford & Swaminathan, 1990). However, little is known about Bayesian procedures in polytomous IRT models.

CHAPTER 4

DESIGN OF THE STUDY AND METHODOLOGY

4.1 Overview of Study

In order to adequately investigate the properties of parameter estimates, a simulation study was conducted. A simulation study is necessary because only by using simulated data is it possible to investigate the accuracy of estimation.

Artificial data were generated for this study according to the generalized partial credit model (GPCM). The generated data was calibrated to obtain MML and Bayesian estimators using the computer programs PARSCALE (Muraki & Bock, 1993). After calibration of the parameter estimates, the properties of the estimates, the accuracy (RMSE), mean squared error (MSE), and bias were examined. In this research, two simulation studies were conducted to study the problem of estimation. In the study I, the properties of MML estimators in the GPCM were examined. In addition, the factors affected parameter estimation in the GPCM were investigated. In the study II, the effectiveness of Bayesian procedures for estimating parameters in the GPCM was investigated and compared the Bayesian procedure with the MML procedure.

4.2 Design of Study

Previous research has indicated that the factors that affect the behavior of parameter estimates of an IRT model are: 1) the characteristics of a test, 2) characteristics of the calibration sample and, 3) characteristics of the estimation procedure.

4.2.1 Test Characteristics

4.2.1.1 Test Length

Test length is an essential factor that influences parameter estimates, because the item response patterns across the items are used in the estimation procedure. For small number of items, it is possible to study all possible response patterns, but for a large number of items to only a sample of the response patterns can be studied. Previous studies with dichotomous IRT models showed test lengths affected the accuracy of the parameter estimates (Stone, 1992; Yen, 1987).

The effect of test lengths was included in this study because previous studies with polytomous model (De Ayala, 1995; Reise and Yu, 1990; Walker-Barnick, 1990) used a fixed test length. In addition, a preliminary study found that test length had an effect on item parameter estimation. That is, item parameter estimates with small number of items (below 10 items) and large number of items (above 30 items) were more accurate than those obtained with a moderate number of items (15 to 25 items). This result may be attributed to the fact that all response patterns can be used with small number items while with more than 15 items an approximation is needed. The approximation works well when the number of items is large but not well when the number of items is small.

A short test (9 items) was included in this study because most of performance assessment tests have a small number of items. Also a moderate (18 items) and a large test (36 items) were also be studied. In each simulation study, the 9 and 18 items were taken from the test with 36 items.

4.2.1.2 The Number of Response Categories

In contrast to binary models, polytomous IRT models contain more item parameters to be estimated because of the additional response categories. The additional number of categories in polytomous IRT models not only result in more parameters to be estimated but also can result in the parameters having extreme values. The additional parameters to be estimated and the extreme value of the parameters seem to have an effect on the accuracy of estimation.

Five response categories for each item were generated in this study because it is the most commonly used number of categories for attitude, achievement, and performance tests. Also three response categories were included in this study to investigate the effect of the number of categories on the accuracy of parameter estimates.

4.2.1.3 Item Parameter Values

An item in a test can be characterized by two item parameters (item difficulty and discrimination parameters). Previous research (Drasgow, 1987; Stone, 1992) with dichotomous IRT models pointed out that the accuracy of estimation depended on the values of the particular item parameters. Polytomous IRT models may have a large item category or step parameter values. It appears that when there are many categories, the parameter estimates for the extreme categories may not be as accurate as that for the middle categories (Reise & Yu, 1990; Choi, Cook, Dodd, 1996).

To make all possible combination of parameter values, item discrimination parameters were classified into three levels; high, middle and low items and item category

(or step) parameter values were classified three levels; easy, moderately, and difficult items. The difficulty levels and discrimination levels were combined to yield items with desired characteristics.

4.2.2 Characteristics of the Calibration Sample

4.2.2.1 Sample Size

In statistical procedures, sample size is a key factor in determining the “quality” of parameter estimates. This is particularly true in complex model such as polytomous IRT models. Prior research on parameter estimation with dichotomous IRT models have shown that sample size is a major factor that affects estimation of item parameters. In polytomous IRT models the interaction of sample size and the number of categories can be expected to affect parameter estimation. When the sample size is small, category may not have a sufficient number of examinees to obtain the accurate parameter estimates.

Reise and Yu (1990) found that at least 500 examinees were needed to achieve an adequate calibration for the 25 test length with five response categories under the GRM. De Ayala (1995) suggested a sample size of ratio 5:1 for the NRM. The actual sample size of the 5:1 ratio in his study was 1000 examinees. Choi, Cook, & Dodd proposed more than 8:1 ration of sample size for the PCM. The actual sample size of ratio 8:1 in their research resulted in more than 500 examinees. To investigate how large a sample size is needed to obtain satisfactory parameter estimates, four different sample sizes; a small (250 examinees), a moderate (500 examinees), and a large (1000 examinees) were examined.

4.2.2.2 Ability Distribution and the Minimum Number of Examinees in Each Category

Ability distribution of examinees is another factor which may affect the quality of parameter estimation. The ability distribution affects the number of responses in each category and in turn affects estimation of parameters. De Ayala (1995) found that as the ability distribution departed from a uniform distribution the accuracy of estimation decreased. He mentioned that the effects of the form of latent distribution on the accuracy (RMSE) might be related to the distribution of responses across item categories. It was found that the uniform distribution produced the greatest dispersion of responses across item categories and that the positively skewed distribution produced least variability in the examinees responses. Inaccuracy of parameter estimates may be related to the insufficient number of examinees across item categories not directly be related to the ability distribution.

In practice, when polytomous scoring is used, the incidence of low frequency categories occurs when this happens. A practitioner has the option of using the data as they exist or of collapsing the low frequency categories into adjacent categories (Brown, 1991). Little is known regarding the effect of insufficient number of examinees in each category on the accuracy of parameter estimates, so it is necessary to study the effect of the minimum number of examinees in each category which occur as a result of the ability distribution on the accuracy of parameter estimates. Unfortunately it is impossible to control the number of examinees in each category for a simulation study. Therefore in this study I ability distributions were included as a factor to examine the tendency of the minimum number of examinees in each category. In the study I four ability distributions

were examined (normal, uniform, positively and negatively skewed distributions). The positively skewed distribution for this study was defined by a χ^2 distribution with twelve degrees of freedom. The negatively skewed distribution was obtained as the mirror image of the positively skewed distribution.

4.2.2.3 Estimation Procedure

Estimation procedure is obviously an essential factor that affects the quality of parameter estimation. Many researchers have studied the effect of estimation procedure on the accurate parameter estimates (Lord, 1986; Mislevy & Stocking, 1989; Swaminathan, 1983; Vale & Gialluca, 1988; Yen, 1987). MMLE with EM is the most popular statistical procedure for obtaining parameter estimates of polytomous IRT models.

MML estimators possess several useful and important properties such as efficiency, consistency and asymptotic normality. The computer program MULTILog (Thissen, 1991) and PARSCALE (Muraki & Bock, 1993) implemented the MMLE procedure with EM algorithm to obtain the parameter estimates for the polytomous IRT models. However there is little known about the properties of the estimators of the polytomous IRT models with the MMLE with EM algorithm.

With MMLE procedure certain data sets can yield unacceptable value of discrimination and difficulty parameter values (Baker, 1992). Bayesian approach may solve this problem by specifying prior information on item parameters. Imposing prior information on item parameters on dichotomous IRT models using Bayes' rule facilitates

estimation with relatively small samples (Harwell & Janosky, 1991; Gifford & Swaminathan, 1990). Computer program MULTLOG (Thissen, 1991) assumes normal prior distributions for item parameters and PARSCALE (Muraki, 1993) assumes log-normal distribution for slope parameter and normal distribution for threshold parameter. Both programs allow users to specify the mean and variance of the prior distributions. Research on the effect of prior distribution on estimation procedure for the polytomous IRT models is another important issue to be explored.

In sum, study I included four factors among those factors described above, sample size (4 levels), test length (3 levels), the number of categories (2 levels), and ability distributions (4 levels). It yielded a four factor design with 96 conditions (Table 1). In the study II, eight different priors for slope and threshold parameters were included as a factor, but ability distributions were not. Prior distributions included in the second study are shown in Table 2. In summary, the factors manipulated in the second study were: Prior distributions (8 levels), Number of categories (2 levels), Test lengths (3 levels), and Sample sizes (3 levels). These four factors in the second study were completely crossed to yield a $8 \times 2 \times 3 \times 4$ factorial design with 192 conditions.

Table 1

Factorial design with 4 factors: 2 x 3 x 4 x 4

# of categories	# of items	ability distributions	sample sizes
3	9	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000
	18	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000
	36	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000
5	9	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000
	18	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000
	36	normal	100
		uniform	250
		positively skewed	500
		negatively skewed	1000

Table 2

Prior distributions for the slope and the threshold parameters

Prior specification				
Type of prior	Slope parameters		Threshold parameters	
	mean	SD	mean	SD
Default 1	1.13	0.6	0.0	2.0
Default 2	1.13	0.6	No prior	
Default 3	No prior		0.0	2.0
True distribution based 1	Mean of the distribution of true slope parameter values	SD of the distribution of true slope parameters	No prior	
True distribution based 2	mean of the distribution of true slope parameter values	default value (0.6)	No prior	
Empirical 1	Polyserial correlation	SD of the distribution of polyserial correlations	No prior	
Empirical 2	Polyserial correlation	default value (0.6)	No prior	
No prior (MMLE)	No prior		No prior	

4.3 Data Generation

Item parameter values for 3 category items used in the study were obtained by analyzing the 1994 NAEP Mathematics test. From the NAEP item parameter estimates, three sets of values were chosen for the slope/discrimination parameters: low discrimination values, less than 0.5; medium discrimination values, between 0.5 and 0.9; high discrimination values, higher than 0.9. Three sets of step difficulty parameters were selected: “easy items” with the step difficulty value for the highest response category less than 0.8; “medium difficulty items” with step difficulty values for the lowest and highest response categories ranging from -3.0 to +3.0 ; “difficult items” with the step difficulty value for the lowest category value higher than 0.8. These discrimination and step difficulty values were crossed to yield nine combinations of “item types”. The nine-item test was constructed with these nine combinations of item parameter values. The eighteen-item test was constructed with two items at each discrimination/difficulty parameter combination; the 36-item test was constructed with four items at each discrimination/difficulty parameter combination. These item parameter values are given Table 3.

Step difficulty parameter values for 5 category items were obtained by adding 0.40 to the last step difficulty parameter value of the 3 category items and by subtracting 0.40 from the first step difficulty parameter value of the 3 category items. The value of 0.40 is the mean difference across items between step difficulty values in 3 category items. The reason for using the same step difficulty parameter values for 5 category items

as 3 category items is to reduce the effect of item parameter values on estimation. Five category item step difficulty parameter values are given Table 4.

Using the item parameter values, item response vectors were generated by randomly sampling θ from the specified distribution and determining the probability of an examinee responding in each category of an item according to the GPCM. For each examinee, cumulative probabilities were obtained for each category. The cumulative probabilities were compared with a random number drawn from a uniform [0,1] distribution. The ordinal position of the first cumulative probability which was greater than the random number was taken as the examinee's response to the item.

For the study II, a negatively skewed distribution was used to generate ability parameter values. This was because study I found that there was no variation in the accuracy of item parameter estimation among normal, positively skewed, negatively skewed, and uniform distributions. The negatively skewed distribution was chosen because it does not reproduce the form of the prior distribution used in PARSCALE. Using the generated ability values, item responses for the GPCM were constructed using the FORTRAN program POLYGEN (Park & Swaminathan, 1996).

It should be noted that although the ability values were drawn from a population with mean zero and standard deviation one, there is no guarantee that the obtained ability values will have a mean of zero and unit standard deviation. Since the ability distribution in PARSCALE is standardized and the item parameter estimates scaled relative to the scale of the ability distribution, to ensure that the item parameter estimates from PARSCALE would be on the same scale as the true item parameters, the generated true

ability values were rescaled to have a mean of 0 and a standard deviation of 1. This obviates the need for equating before making comparisons between the true values and estimates.

In studying the effect of test length on estimation, the tests were lengthened systematically, i.e., tests were lengthened by adding items to the original set. The nine item test is a subtest of the 18 item test which in turn is a subtest of the thirty-six item test. The same principle was used in generating examinee trait values. That is, as sample size was increased, the same examinees as in the smaller data set, along with additional examinees, were administered the test. The purpose of generating the data in this way was to minimize the variability due to sampling from the population of true values, and thereby to facilitate interpretation of trends in the results.

4.4 Criteria for Evaluating Adequacy of the Estimates

The criteria used to evaluate the Bayes and the marginal maximum likelihood estimators of the polytomous IRT models were accuracy of estimates, sampling variance of estimates, and bias of estimates over replications. Accuracy of parameter estimates is measured by the mean squared difference (MSD) between the true and estimated parameter. The smaller the MSD, the more accurate the estimates. The MSD can be separated into the variance of the estimates over replications (VAR) and Squared Bias, defined as the squared difference between the true parameter value and the mean of the estimates over replications (Gifford & Swaminathan, 1990). When r replications are carried out,

$$\frac{\sum_{k=1}^r (\hat{T}_{ik} - T_i)^2}{r} = \frac{\sum_{k=1}^r (\hat{T}_{ik} - \bar{\hat{T}}_i)^2}{r} + (\bar{\hat{T}}_i - T_i)^2 \quad (25)$$

where \hat{T}_i is the estimate of the parameter T_i in the r -th replication. The term on the left of the equation is the MSD, while the terms on the right are variance and squared bias, respectively. This decomposition of MSD into sampling variance and squared bias permits the identification of the causes of errors in estimation. Ideally, bias should be zero, in which case the accuracy of the parameter estimates is determined solely by the variance of the estimates. On the other hand, if the variance is small, then bias is the main cause of the error in estimation. Without replications and this decomposition, the above determination cannot be made. The above decomposition permits the study of MSD, variance, and bias at the item level, or at the test level by averaging the quantities computed in the above manner over the items. More important, this decomposition permits grouping items according to item types and examining the reasons for poor estimation of parameters.

In order to summarize the information, MSD, variance, and bias are averaged across the items. In this study, the square root of each of these quantities is reported. RMSE (root mean squared error) is the square root of MSD and is used as an index of the accuracy of the parameter estimates. To obtain the average RMSE across items, the

MSDs for the items are averaged and the square root of the average is taken. The standard deviation of the estimates (square root of VAR) is used as an index of the variability of estimates. The average standard deviation across items is obtained by averaging the variance, and then taking the square root. Bias is the square root of the squared bias quantity described above. This definition of bias corresponds to the conventional definition of the term bias in that the mean of the estimates and the true value of the parameter are compared; while taking the square root of the squared term removes the sign from the bias indicator, the magnitude of the mean compared to the true value provides information as to whether the parameter is being over- or under-estimated . The bias was averaged over categories and items in the same manner as MSD and variance. These indices were then subject to descriptive statistical analysis.

In evaluating the effect of increasing the number of categories on parameter estimation, the MSD, variances, and bias indices were averaged across the step difficulty parameters to yield a single MSD value for the category parameters for each item. As mentioned earlier, the step difficulty parameters, b_1 and b_2 in the three-category items were kept the same as the step difficulty parameters b_2 and b_3 in the five-category items. The accuracy of estimation of these parameters when the number of categories changed can be compared directly. However, the category parameters b_1 and b_4 in the five-category items do not have any counterparts in the three- category items. To avoid any inconsistency, the MSD, Variance, and bias indices were averaged over the category parameters. In increasing the number of categories from three to five, the slope

parameter was kept at the same value, and hence the effect of increasing the number of categories on the slope parameter can be assessed directly.

In addition to the descriptive analysis, the effect of each factor on the accuracy, the variance, and the bias of the item parameter estimates in the GPCM were determined using analysis of variance procedures. The dependent variables were the RMSE, variance, and bias of the estimates of the slope and the step parameter estimates. The 4 factors (the number of categories, ability distributions, test lengths, and sample sizes) were used as the independent variables in the analysis for study I. Prior distributions were included in the independent variables instead of ability distributions in the analysis for study II. The purpose of the analysis of variance is to determine, in a descriptive sense, which factors influenced the outcome variables, RMSE, standard deviation, and bias. Given this, interpretation of the levels of significance of the statistical tests was not of primary interest. In analyzing the data, separate univariate analyses rather than a multivariate analyses were performed, partly because of the inherent "almost" linear dependencies among these dependent variables and also because of the descriptive emphasis on the analyses. In addition, to keep the analyses tractable, the interaction terms were suppressed.

4.5 Calibration

The generated response data set were calibrated according to the GPCM using the computer program PARSCALE (Muraki, 1993). PARSCALE can be used for parameters

of dichotomous IRT models, ordered polytomous models : the GRM, the PCM, the GPCM and the RSM, but not the nominal models.

Although item and ability parameters are theoretically invariant in item response models, there is a basic indeterminacy in the model when both ability and item parameters are unknown. In order to anchor the scale and to provide a unique solution, it is necessary in most estimation procedures to fix location by setting the mean of the ability distribution to 0 and to fix scale by setting the standard deviation of the distribution to 1. To put the estimates from computer program to the same metric of the true item parameters, the generated ability values were standardized to have a mean of 0 and a standard deviation of 1. With this standardization the item parameter estimates were on the same scale as parameters and the estimates and parameters could be compared.

Table 3

True item parameter values for 3 category items

ID	a	b1	b2
1	0.392	-1.201	-0.199
2	0.379	-1.205	1.429
3	0.438	0.854	1.269
4	0.788	-0.907	0.437
5	0.718	0.405	1.101
6	0.68	0.902	1.802
7	1.063	-0.06	0.628
8	1.139	-0.113	1.303
9	0.995	0.824	1.248
10	0.394	0.29	0.738
11	0.493	-1.109	0.897
12	0.385	0.973	1.349
13	0.635	-1.066	-0.288
14	0.855	0.088	1.006
15	0.61	0.901	1.775
16	0.958	0.069	0.303
17	0.922	-0.625	1.341
18	1.058	0.801	1.508
19	0.472	-1.143	0.629
20	0.328	-1.143	1.573
21	0.455	0.842	1.456
22	0.716	-0.306	0.466
23	0.68	0.251	1.079
24	0.571	0.802	1.701
25	0.989	-0.145	0.307
26	1.054	0.656	1.058
27	1.174	0.809	1.263
28	0.489	-0.574	-0.286
29	0.436	-0.229	1.137
30	0.411	0.984	1.39
31	0.537	-0.574	-0.286
32	0.801	0.328	0.942
33	0.684	0.801	1.401
34	1.001	0.438	0.772
35	1.201	0.124	1.342
36	0.916	0.801	1.239

Table 4

True item parameter values for 5 category items

ID	a	b1	b2	b3	b4
1	0.392	-1.601	-1.201	-0.199	0.201
2	0.379	-1.605	-1.205	1.429	1.829
3	0.438	0.454	0.854	1.269	1.669
4	0.788	-1.307	-0.907	0.437	0.837
5	0.718	0.005	0.405	1.101	1.501
6	0.68	0.502	0.902	1.802	2.202
7	1.063	-0.46	-0.06	0.628	1.028
8	1.139	-0.513	-0.113	1.303	1.703
9	0.995	0.424	0.824	1.248	1.648
10	0.394	-0.11	0.29	0.738	1.138
11	0.493	-1.509	-1.109	0.897	1.297
12	0.385	0.573	0.973	1.349	1.749
13	0.635	-1.466	-1.066	-0.288	0.112
14	0.855	-0.312	0.088	1.006	1.406
15	0.61	0.501	0.901	1.775	2.175
16	0.958	-0.331	0.069	0.303	0.703
17	0.922	-1.025	-0.625	1.341	1.741
18	1.058	0.401	0.801	1.508	1.908
19	0.472	-1.543	-1.143	0.629	1.029
20	0.328	-1.543	-1.143	1.573	1.973
21	0.455	0.442	0.842	1.456	1.856
22	0.716	-0.706	-0.306	0.466	0.866
23	0.68	-0.149	0.251	1.079	1.479
24	0.571	0.402	0.802	1.701	2.101
25	0.989	-0.545	-0.145	0.307	0.707
26	1.054	0.256	0.656	1.058	1.458
27	1.174	0.409	0.809	1.263	1.663
28	0.489	-0.974	-0.574	-0.286	0.114
29	0.436	-0.629	-0.229	1.137	1.537
30	0.411	0.584	0.984	1.39	1.79
31	0.537	-0.974	-0.574	-0.286	0.114
32	0.801	-0.072	0.328	0.942	1.342
33	0.684	0.401	0.801	1.401	1.801
34	1.001	0.038	0.438	0.772	1.172
35	1.201	-0.276	0.124	1.342	1.742
36	0.916	0.401	0.801	1.239	1.639

CHAPTER 5

RESULTS

5.1 Introduction

Two simulation studies were carried out for the investigation of the properties of marginal maximum likelihood and Bayesian estimators in the Generalized Partial Credit model (GPCM). Marginal maximum likelihood (MML) and Bayesian estimates in the GPCM, as obtained through the computer program PARSCALE, were compared with respect to accuracy, variance and bias. In addition, the effectiveness of Bayesian estimates with respect to the specification of priors was investigated. In this chapter, the results of study I that focused on the properties of marginal maximum likelihood estimators in the GPCM are presented first. The results of study II that examined the effectiveness of Bayesian procedures for estimating parameters in the GPCM are presented next.

5.2 Results of Study I

5.2.1 Accuracy of Estimation

The average RMSE over 100 replications across all conditions for 3- and 5-category items is reported in Table 5.

Table 5

The average RMSE across all conditions for 3 and 5 category items

Distribution	Category	Sample size	RMSE for the slope parameters			mean RMSE for category parameters			
			Test length			Test length			
			9	18	36	9	18	36	
normal	3	100	0.276	0.170	0.114	0.373	0.322	0.181	
		250	0.145	0.095	0.071	0.219	0.158	0.126	
		500	0.100	0.063	0.045	0.150	0.116	0.105	
		1000	0.071	0.045	0.032	0.110	0.089	0.092	
	5	100	0.212	0.152	0.114	0.408	0.293	0.215	
		250	0.122	0.077	0.063	0.264	0.188	0.150	
		500	0.084	0.055	0.055	0.190	0.138	0.125	
		1000	0.063	0.045	0.045	0.141	0.114	0.112	
	Uniform	3	100	0.300	0.155	0.114	0.440	0.281	0.184
			250	0.164	0.095	0.071	0.224	0.163	0.122
			500	0.126	0.071	0.055	0.165	0.124	0.102
			1000	0.095	0.055	0.045	0.134	0.102	0.100
5		100	0.226	0.134	0.118	0.422	0.286	0.206	
		250	0.130	0.089	0.071	0.269	0.182	0.145	
		500	0.100	0.063	0.055	0.198	0.142	0.129	
		1000	0.063	0.055	0.045	0.141	0.120	0.118	
positive		3	100	0.253	0.148	0.105	0.352	0.247	0.186
			250	0.155	0.095	0.063	0.216	0.152	0.124
			500	0.104	0.071	0.045	0.152	0.116	0.100
			1000	0.084	0.045	0.032	0.109	0.089	0.089
	5	100	0.249	0.138	0.114	0.421	0.300	0.220	
		250	0.158	0.100	0.063	0.266	0.194	0.148	
		500	0.089	0.055	0.045	0.191	0.147	0.125	
		1000	0.063	0.045	0.045	0.141	0.117	0.112	
	negative	3	100	0.266	0.141	0.100	0.453	0.274	0.202
			250	0.138	0.084	0.055	0.253	0.163	0.120
			500	0.105	0.063	0.045	0.177	0.114	0.097
			1000	0.077	0.045	0.032	0.124	0.087	0.084
5		100	0.200	0.161	0.100	0.450	0.341	0.212	
		250	0.118	0.071	0.055	0.282	0.184	0.136	
		500	0.084	0.055	0.045	0.201	0.135	0.109	
		1000	0.063	0.032	0.032	0.147	0.102	0.092	

The ANOVA result of the average RMSE (averaged over estimates of the slope and category difficulty parameters across all conditions) is reported in Table 6.

Table 6

Results of ANOVA for Root Mean Squared Error (RMSE)

Source	df	Parameter	F value	P value
Number of Categories	1	slope	2.84	.096
	1	category parameter	11.95	.001
Distribution	3	slope	.78	.506
	3	category parameter	.52	.668
Test length	2	slope	83.37	.000
	2	category parameter	83.37	.000
Sample size	3	slope	121.31	.000
	3	category parameter	147.43	.000

The result of the ANOVA of RMSE in Table 6 show that test lengths and sample sizes influence the accuracy of estimates of the slope and the category parameters. The ability distribution does not seem to have an effect on the accuracy of estimation of the parameters. The number of categories in each item influences the accuracy of estimation of the category parameters but not the accuracy of estimation of the slope parameters.

Figures 3 and 4 (a) provide graphical description of the effect of various factors on the average RMSE of the estimates of item parameters for 3- and 5-category items across sample sizes (100, 250, 500 and 1000) and test lengths (9, 18, and 36 items) based on the normal ability distribution.

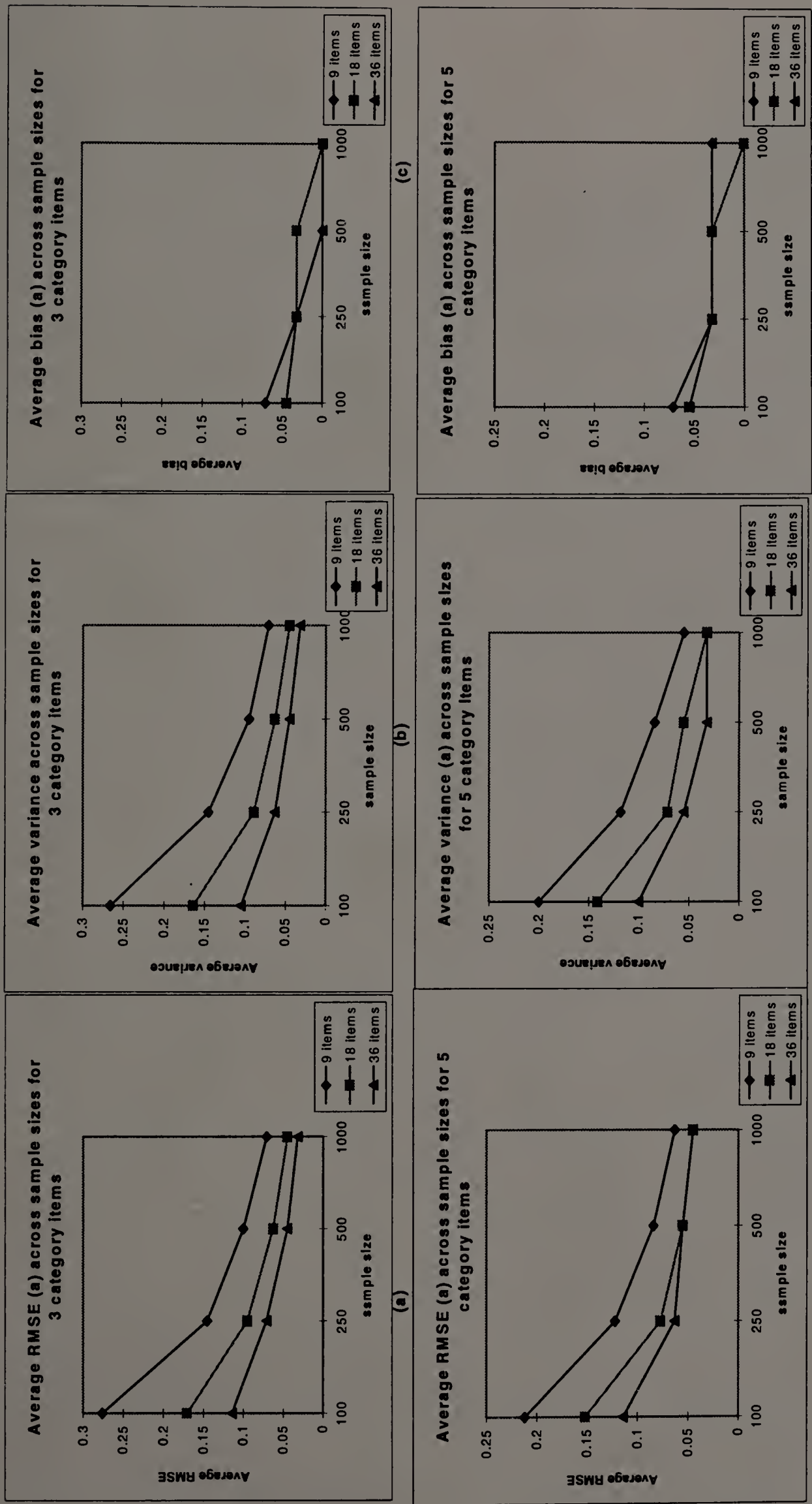


Figure 3. Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on normal distribution

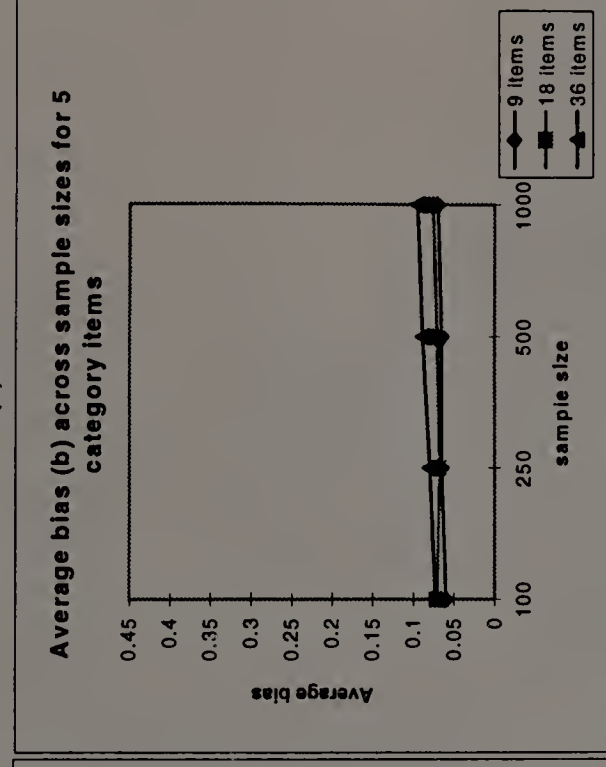
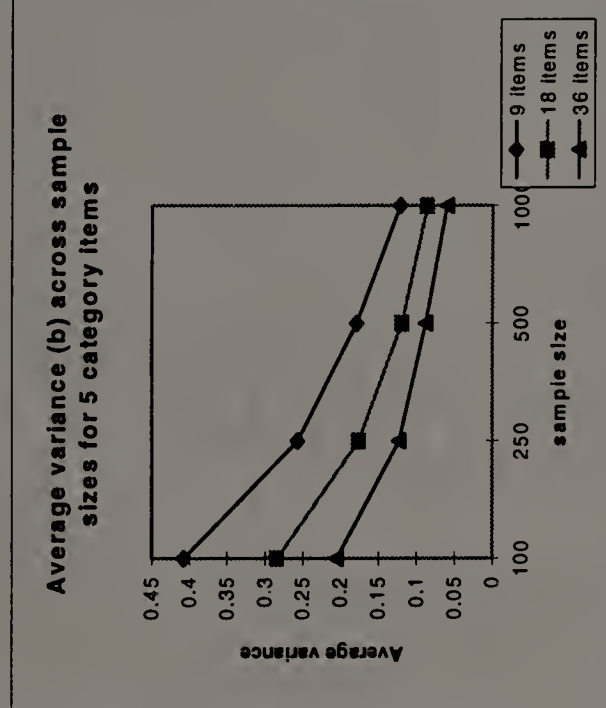
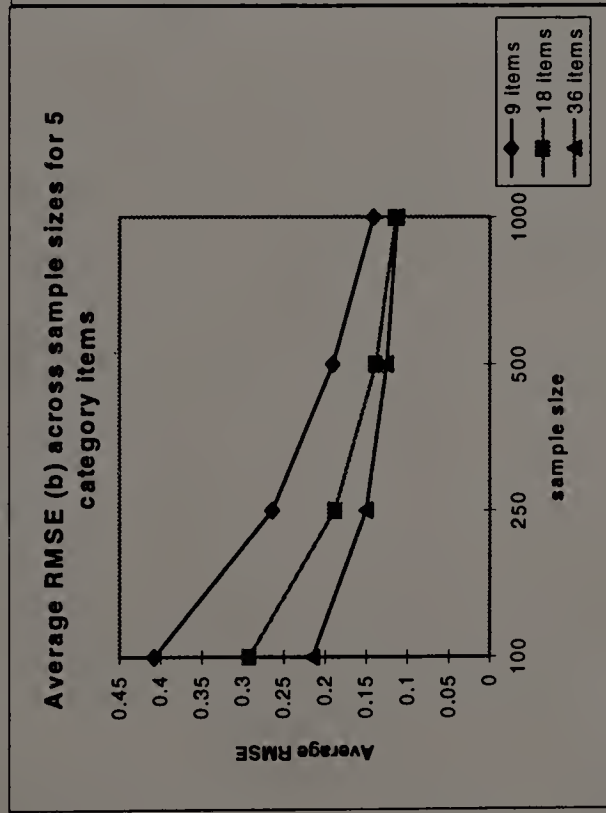
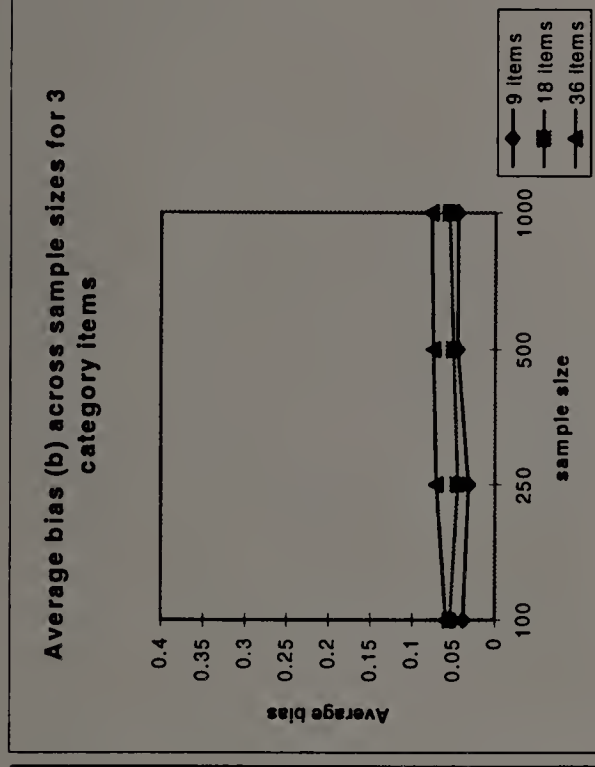
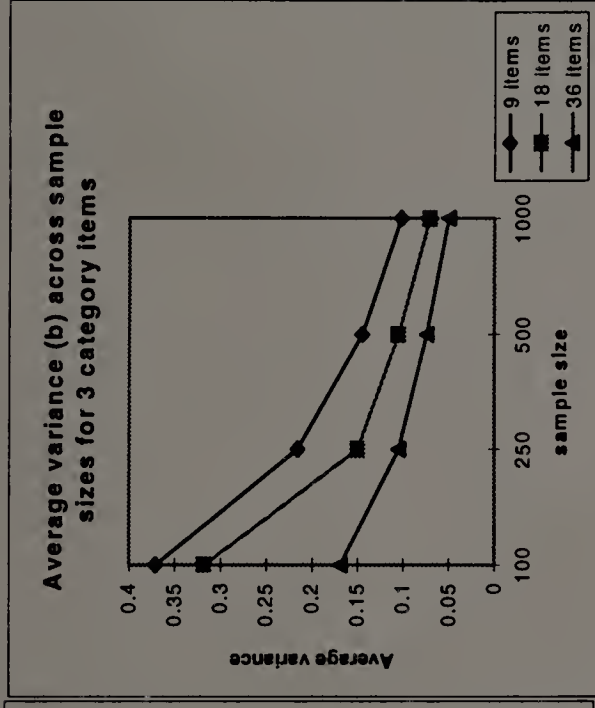
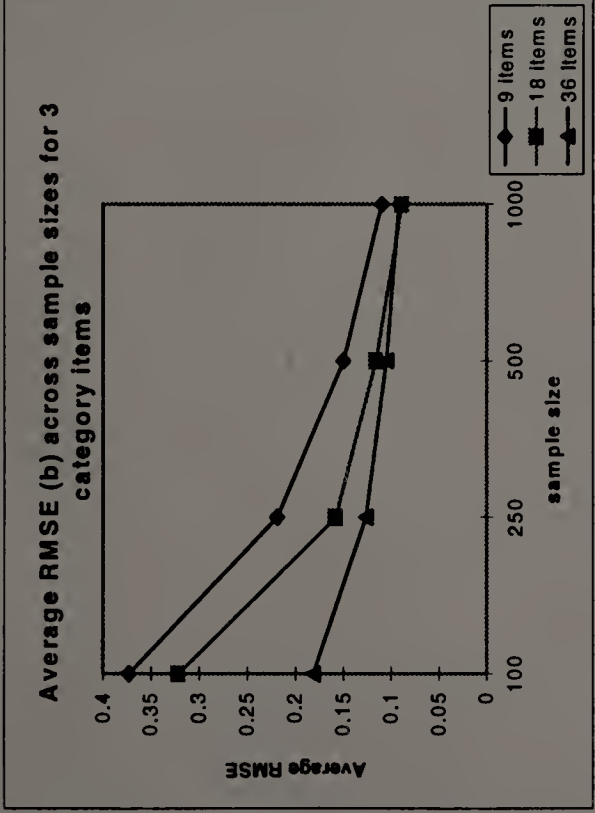


Figure 4. Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on normal distribution

Figures based on other than normal distributions are displayed in the Appendix A, since the results of the ANOVA of RMSE showed that the ability distributions did not affect the accuracy of estimation.

Figure 3 (a) is for the RMSE of the estimates of slope parameters. The upper figure is for the 3-category items and the lower one is for the 5-category items. Figure 4 (a) is for the mean RMSE of the estimates of category difficulty parameters; again, the upper figure is for the 3- category items and the lower one is for the 5-category items. Figures 3 and 4 (a) confirm the ANOVA finding that sample size and test length have an effect on the accuracy of estimates of both the slope and category parameters. The average RMSE decreased as test length increased from 9 to 36 items. The most noticeable decrease in the RMSE of estimates of both slope and category parameters occurred when test length increased from 9 to 18 items.

As expected, the accuracy of estimates increased (as shown by decreasing RMSE) as sample size increased. The greatest improvement in accuracy of the estimates of the slope and category parameters occurred when sample size increased from 100 to 250. The improvement beyond a sample size of 250 was modest.

Increasing the number of categories from three to five had a negative effect on the estimation of category parameters. The results of the ANOVA showed that this effect was significant. It appears that, for a fixed sample size, as the number of categories increased, the number of examinees falling in a category decreased. On the other hand, the effect of increasing the number of categories had a modest positive effect, albeit not statistically significant, on the estimation of the slope parameter. It can be conjectured that since the slope parameter is constant across the categories for a give item, increasing

the number of categories provided more stable information regarding the slope parameter and hence resulted in better estimation of the parameter. The modest improvement obtained by going from a three-category item to a five-category item vanished as the sample size and test length increased, as shown in Figures 3 and 4(a).

5.2.2 Variance and Bias

In addition to the accuracy of estimates, the variance and bias of the estimates are important quantities in evaluating the quality of parameter estimation. The source of the difference between the estimates and the true parameter values, that is the accuracy of estimates, can be partitioned into sampling error, variance, and systematic bias. The sampling error is, in reality, is the square of the standard error of the estimate obtained empirically. If an estimator shows great variation over repeated samples, i.e., has a large standard error, then the parameter will be estimated with less accuracy.

The average variance over replications across all conditions is reported in Table 7. The results of the ANOVA of the average variance of the estimates of the slope and the mean of the category parameters are reported in Table 8. Table 8 shows that the number of categories in each item, test lengths, and sample sizes have an impact on the variance of estimates of the slope and the category difficulty parameters, but the distribution from which the samples are drawn does not affect the estimates with respect to variance.

Table 7

The average variance across all conditions for 3 and 5 category items

Distribution	Category	Sample size	Variance for slope parameters			Mean variance for category parameters		
			Test length			Test length		
			9	18	36	9	18	36
normal	3	100	0.266	0.164	0.105	0.371	0.318	0.169
		250	0.145	0.089	0.063	0.216	0.150	0.105
		500	0.095	0.063	0.045	0.145	0.105	0.074
		1000	0.071	0.045	0.032	0.102	0.071	0.050
	5	100	0.200	0.141	0.100	0.408	0.284	0.205
		250	0.118	0.071	0.055	0.257	0.176	0.123
		500	0.084	0.055	0.032	0.179	0.119	0.088
		1000	0.055	0.032	0.032	0.122	0.086	0.060
Uniform	3	100	0.270	0.141	0.105	0.413	0.271	0.171
		250	0.141	0.084	0.055	0.211	0.145	0.095
		500	0.100	0.055	0.045	0.143	0.095	0.067
		1000	0.071	0.045	0.032	0.102	0.067	0.045
	5	100	0.205	0.118	0.100	0.418	0.280	0.190
		250	0.114	0.077	0.055	0.254	0.140	0.114
		500	0.084	0.055	0.032	0.174	0.113	0.083
		1000	0.055	0.032	0.032	0.129	0.081	0.056
positive	3	100	0.239	0.141	0.095	0.349	0.244	0.173
		250	0.145	0.089	0.055	0.212	0.143	0.100
		500	0.095	0.063	0.045	0.145	0.105	0.071
		1000	0.071	0.045	0.032	0.100	0.071	0.045
	5	100	0.235	0.126	0.095	0.416	0.294	0.202
		250	0.152	0.089	0.055	0.257	0.180	0.120
		500	0.084	0.055	0.032	0.173	0.124	0.083
		1000	0.063	0.032	0.032	0.118	0.084	0.056
negative	3	100	0.266	0.145	0.105	0.438	0.270	0.196
		250	0.130	0.084	0.055	0.244	0.157	0.107
		500	0.095	0.063	0.045	0.164	0.105	0.077
		1000	0.063	0.045	0.032	0.112	0.074	0.055
	5	100	0.197	0.071	0.089	0.449	0.223	0.205
		250	0.114	0.071	0.045	0.281	0.179	0.122
		500	0.077	0.055	0.032	0.187	0.125	0.088
		1000	0.055	0.032	0.032	0.131	0.089	0.058

Table 8

Results of ANOVA for variance

Source	df	Parameter	F value	P value
Number of Categories	1	slope	7.55	.007
	1	category parameter	10.19	.002
Distribution	3	slope	.16	.921
	3	category parameter	.81	.491
Test length	2	slope	32.38	.000
	2	category parameter	141.96	.000
Sample size	3	slope	52.10	.000
	3	category parameter	224.74	.000

Figures 3 and 4 (b) provide summaries of the average variance for the item parameters for for 3 and 5 category items across sample sizes and test lengths. These figures are for the normal distribution of ability since, as indicated above, the ability distribution had no effect on the variance of the estimates. Figure 3 and 4 (b) reveal that the pattern of results for the variance is almost identical to that of RMSE. Sample size and test length have a clear effect on the variance of the estimates slope and category parameters. As test length and sample size increased, the variance decreased along with RMSE. The decrease in variance is most noticeable when the number of items increased from 9 to 18 items and sample size increase from 100 to 250. In addition, the average variance of the slope parameters decreased, but that of category difficulty parameters increased as the number of categories in each item.

If an estimator is unbiased, the mean of the estimates will converge to the true value as the number of replications approaches infinity. Consequently, the difference between the

estimate and the true parameter value, MSD, is attributable to sampling error, or variance of the estimates. The average bias, over replications across all conditions is reported in Table 10.

The result of the ANOVA for the average bias over all conditions for the estimates of the slope and the mean of category difficulty parameters is reported in Table 9.

Table 9
Result of ANOVA for bias

Source	df	Estimates	F value	P value
Category	1	slope	.81	.369
	1	category parameter	15.90	.000
Distribution	3	slope	4.36	.007
	3	category parameter	2.95	.037
Test length	2	slope	4.86	.010
	2	category parameter	1.98	.144
Sample size	3	slope	10.93	.000
	3	category parameter	.31	.821

Table 10 shows that test length and sample size influenced the bias in the estimates of the slope parameter, but not that of the category parameters, i.e., as the sample size increased, the bias in the estimates of the slope parameters changed while that in the category parameters did not. This implies that the estimators of the category parameters may not only be biased but also may not be consistent. The number of categories in each item affected the bias in the estimates of the category parameters, but

not that in the slope parameter. True ability distributions had an impact on the average bias in the estimates of both item parameters.

Figures 5 and 6 along with Figures 3 and 4 display the trends with respect to bias. Uniform ability distribution produced the largest bias for the slope parameter (.053) and the category parameters (0.081). Negatively skewed distribution produced the smallest bias for the slope parameter (.027). Most importantly, the average bias in the slope parameters decreased as the sample size and test length increased. However, the bias in the category parameters remained constant as sample size and test length increased with the amount of bias increasing as the number of categories increased.

5.3 Results of Study II

5.3.1 Accuracy of Estimation

The average RMSE over 100 replications across all conditions for 3 and 5 category items is reported in Table 11 and Table 12. Default priors for threshold parameters did not result in convergence with small sample sizes (100 and 250 examinees) across all test lengths while specification of default priors for both slope and threshold parameters did not result in convergence with 100 examinees in the nine-item test with 3 categories.

Table 10

The average bias across all conditions for 3 and 5 category items

Distribution	Category	Sample size	Bias for slope parameters			Mean Bias for category parameters			
			Test length			Test length			
			9	18	36	9	18	36	
normal	3	100	0.071	0.045	0.045	0.038	0.054	0.059	
		250	0.032	0.032	0.032	0.032	0.045	0.071	
		500	0.000	0.032	0.000	0.045	0.050	0.074	
		1000	0.000	0.000	0.000	0.045	0.055	0.077	
	5	100	0.071	0.055	0.055	0.060	0.071	0.073	
		250	0.032	0.032	0.032	0.065	0.068	0.080	
		500	0.032	0.032	0.032	0.065	0.071	0.089	
		1000	0.032	0.000	0.032	0.070	0.076	0.095	
	Uniform	3	100	0.130	0.063	0.045	0.069	0.063	0.063
			250	0.084	0.045	0.032	0.079	0.074	0.077
			500	0.077	0.045	0.032	0.086	0.077	0.083
			1000	0.071	0.032	0.032	0.086	0.077	0.089
5		100	0.100	0.063	0.063	0.089	0.068	0.082	
		250	0.063	0.045	0.045	0.091	0.083	0.091	
		500	0.055	0.032	0.045	0.096	0.085	0.098	
		1000	0.000	0.032	0.045	0.058	0.088	0.103	
positive		3	100	0.084	0.045	0.045	0.038	0.038	0.067
			250	0.055	0.032	0.032	0.027	0.055	0.074
			500	0.045	0.032	0.000	0.027	0.050	0.074
			1000	0.045	0.000	0.000	0.043	0.054	0.074
	5	100	0.084	0.055	0.055	0.067	0.060	0.087	
		250	0.045	0.032	0.032	0.076	0.075	0.088	
		500	0.032	0.032	0.032	0.075	0.079	0.095	
		1000	0.032	0.032	0.032	0.071	0.077	0.097	
	negative	3	100	0.032	0.000	0.032	0.115	0.050	0.050
			250	0.045	0.000	0.000	0.069	0.045	0.055
			500	0.045	0.032	0.000	0.067	0.045	0.059
			1000	0.055	0.032	0.000	0.054	0.045	0.063
5		100	0.000	0.045	0.045	0.081	0.049	0.059	
		250	0.032	0.000	0.000	0.070	0.052	0.059	
		500	0.032	0.000	0.000	0.069	0.047	0.067	
		1000	0.032	0.000	0.000	0.068	0.053	0.070	

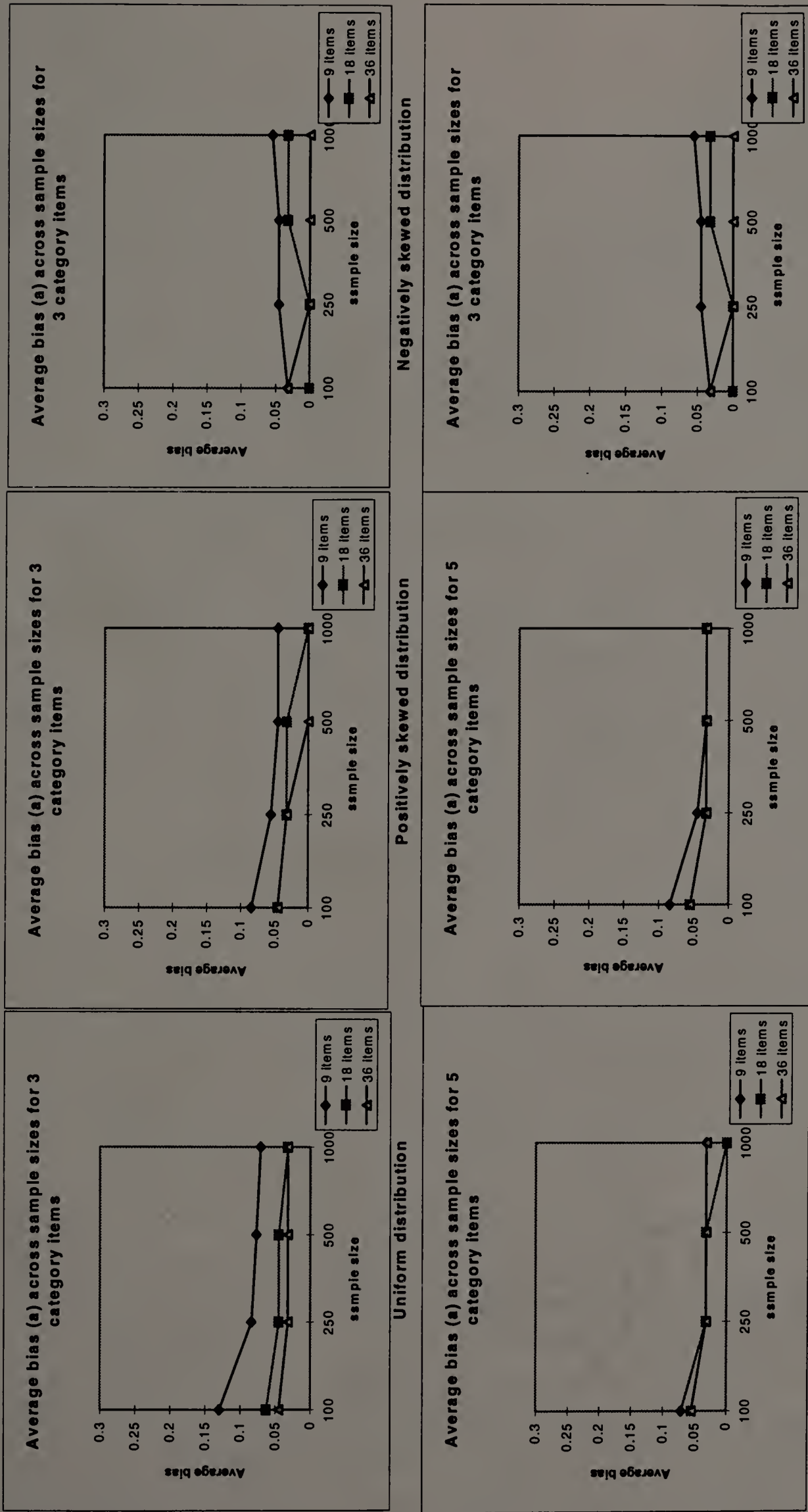


Figure 5. Average bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items

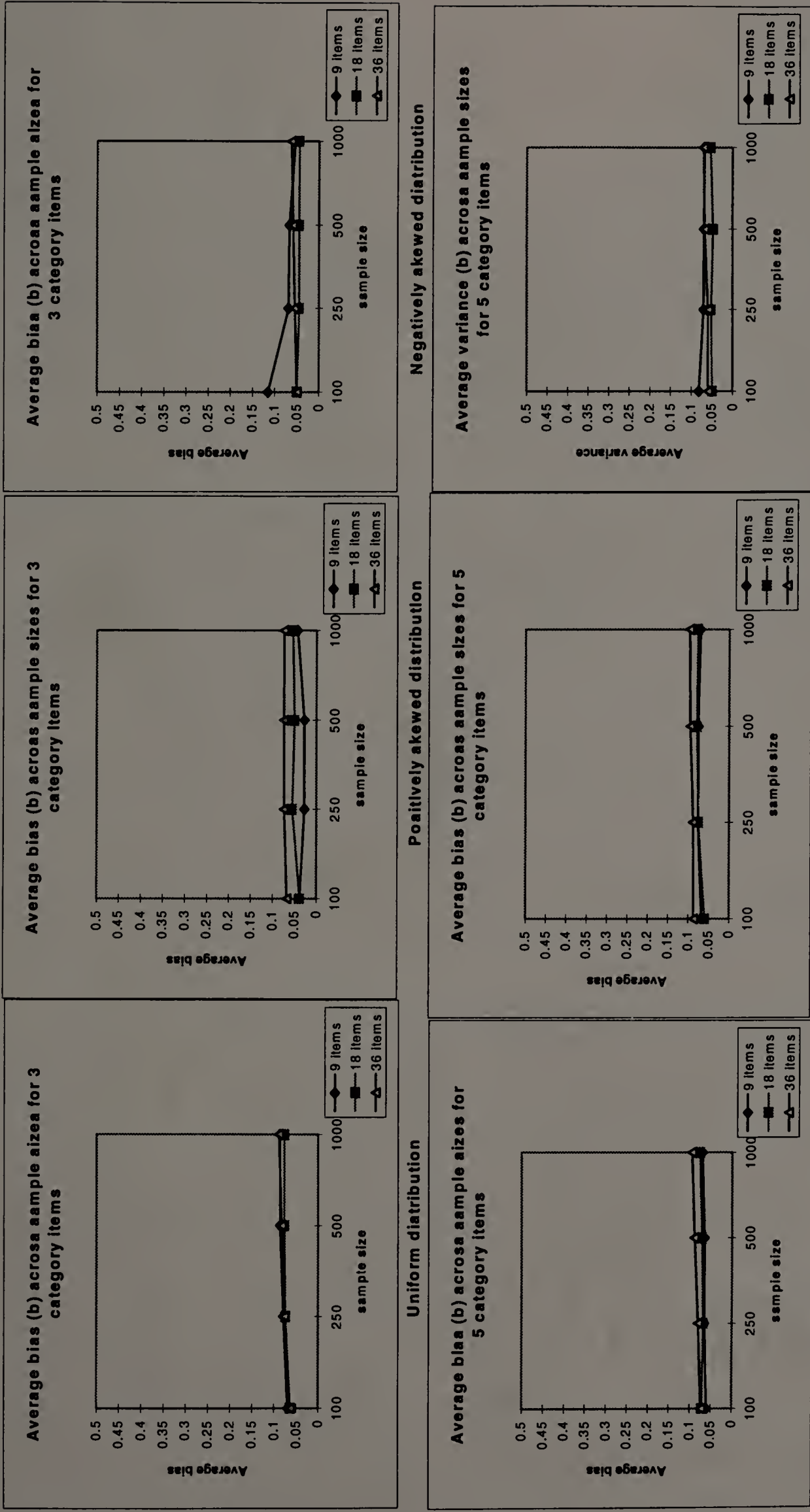


Figure 6. Average bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items

Table 11

Average RMSE of estimates of slope parameters across different priors
for 3 and 5 category items

		MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default	
3 category	9 Items	100	0.268	0.206	0.168	0.178	0.317	0.297	*****	*****
		250	0.138	0.127	0.128	0.129	0.136	0.122	0.127	*****
		500	0.105	0.096	0.102	0.102	0.093	0.094	0.096	0.105
		1000	0.077	0.075	0.081	0.08	0.065	0.072	0.075	0.08
	18 Items	100	0.145	0.15	0.108	0.112	0.207	0.176	0.15	*****
		250	0.084	0.086	0.077	0.077	0.077	0.078	0.085	*****
		500	0.063	0.061	0.064	0.063	0.065	0.058	0.061	0.067
		1000	0.045	0.045	0.049	0.049	0.042	0.044	0.045	0.049
	36 Items	100	0.11	0.132	0.072	0.078	0.101	0.108	0.132	*****
		250	0.063	0.078	0.05	0.052	0.074	0.067	0.077	*****
		500	0.045	0.048	0.037	0.037	0.047	0.04	0.048	0.039
		1000	0.032	0.031	0.027	0.027	0.032	0.028	0.031	0.027
5 category	9 Items	100	0.2	0.182	0.15	0.155	0.324	0.31	0.180	*****
		250	0.118	0.111	0.108	0.109	0.198	0.153	0.111	0.116
		500	0.084	0.08	0.083	0.083	0.112	0.085	0.080	0.085
		1000	0.063	0.059	0.063	0.063	0.064	0.055	0.059	0.063
	18 Items	100	0.127	0.148	0.099	0.101	0.289	0.27	0.142	*****
		250	0.071	0.083	0.065	0.066	0.135	0.101	0.082	0.071
		500	0.055	0.056	0.050	0.051	0.096	0.071	0.056	0.052
		1000	0.032	0.037	0.036	0.036	0.062	0.043	0.037	0.037
	36 Items	100	0.1	0.133	0.072	0.075	0.215	0.194	0.137	*****
		250	0.055	0.083	0.043	0.047	0.126	0.123	0.083	0.054
		500	0.045	0.057	0.033	0.036	0.094	0.073	0.057	0.04
		1000	0.032	0.041	0.027	0.029	0.078	0.058	0.041	0.031

Table 12

Average RMSE of estimates of step difficulty parameters across different priors
for 3 and 5 category items

		MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default	
3 category	9 items	100	0.445	0.302	0.312	0.324	0.337	0.336	*****	*****
		250	0.253	0.216	0.219	0.226	0.225	0.224	0.218	*****
		500	0.179	0.158	0.159	0.164	0.158	0.164	0.159	0.184
		1000	0.126	0.118	0.119	0.121	0.117	0.12	0.118	0.126
	18 items	100	0.267	0.225	0.214	0.221	0.241	0.236	0.239	*****
		250	0.17	0.157	0.147	0.15	0.157	0.157	0.155	*****
		500	0.128	0.117	0.111	0.112	0.128	0.115	0.115	0.122
		1000	0.105	0.088	0.085	0.084	0.096	0.087	0.088	0.084
	36 items	100	0.204	0.195	0.163	0.166	0.192	0.184	0.194	*****
		250	0.141	0.147	0.119	0.119	0.146	0.134	0.142	*****
		500	0.124	0.118	0.099	0.098	0.118	0.104	0.115	0.09
		1000	0.116	0.097	0.086	0.084	0.1	0.089	0.097	0.08
5 category	9 items	100	0.45	0.384	0.389	0.39	0.396	0.399	0.380	*****
		250	0.282	0.259	0.268	0.272	0.268	0.258	0.259	0.283
		500	0.201	0.191	0.194	0.196	0.193	0.189	0.191	0.2
		1000	0.147	0.144	0.146	0.146	0.144	0.142	0.144	0.147
	18 items	100	0.341	0.27	0.282	0.286	0.291	0.289	0.263	*****
		250	0.184	0.182	0.181	0.181	0.201	0.187	0.181	0.185
		500	0.135	0.137	0.135	0.135	0.159	0.145	0.137	0.135
		1000	0.102	0.106	0.103	0.104	0.124	0.111	0.106	0.103
	36 items	100	0.212	0.207	0.19	0.197	0.233	0.227	0.198	*****
		250	0.136	0.148	0.133	0.133	0.166	0.164	0.148	0.135
		500	0.109	0.121	0.107	0.108	0.143	0.13	0.121	0.109
		1000	0.092	0.101	0.091	0.091	0.127	0.112	0.101	0.091

The ANOVA results for the average RMSE (averaged over estimates of the slope and category difficulty parameters across all conditions) is reported in Table 13. The results show that test length, sample size, and prior distribution influence the accuracy of estimates of the slope and the step difficulty parameters. The number of categories in each item influences the accuracy of estimation of the step difficulty parameters but not the accuracy of estimation of the slope parameters.

Table 13
Results of ANOVA for RMSE

Source	df	Estimates	F value	P value
Number of categories	1	Slope	.68	.411
	1	Step difficulty parameters	36.55	.000
Test length	2	Slope	63.93	.000
	2	Step difficulty parameters	170.51	.000
Sample size	3	Slope	139.28	.000
	3	Step difficulty parameters	286.43	.000
Prior distributions	6	Slope	10.78	.000
	6	Step difficulty parameters	2.27	.039

Graphical descriptions of the effect of various factors on the average RMSE of the estimates of item parameters for 3- and 5-category items across sample sizes (100, 250, 500 and 1000) and different prior distributions (no prior, default priors, empirical priors, true distribution-based priors) for each test length are provided in Figure 7 for the slope parameter and in Figure 8 for the step difficulty parameters. Only the results from six priors (No Prior, Default 2, Empirical 1 and 2, and True Distribution-based 1 and 2) are

given in Figures 7 and 8 because Default 3 prior did not result in convergence for small sample sizes and short tests. Default 1 prior was also omitted because it produced almost identical results to those of Default prior 2.

Figures 7 and 8 confirm the ANOVA finding that sample sizes and test length have an effect on the accuracy of estimates of both slope and step difficulty parameters. The average RMSE decreases as test lengths increase from 9 to 36 items. The most noticeable decrease in RMSE of estimates of both types of item parameters appeared when test length increased from 9 to 18 items. As expected, accuracy of estimates increased (as shown by decreasing RMSE) as sample size increased. The greatest improvement in accuracy of the estimates of slope and category parameters occurred when sample size increased from 100 to 250. The improvement beyond a sample size of 250 was modest.

Prior distributions differed with respect to their effects on the accuracy of estimation in small samples while in large samples, their effects were reduced. In using the true distribution of parameters to specify priors, the *mean* of the true distribution was used for as the mean of the prior distribution for *all* the slope parameters. This specification will result in exact priors only for those parameters whose values agree with the mean of the distribution; for other parameters, this prior specification will result in incorrect priors. Using the mean of the true distribution as the mean of the prior distribution for all the parameters thus permits the examination of results when the prior specification is correct and also when it is incorrect.

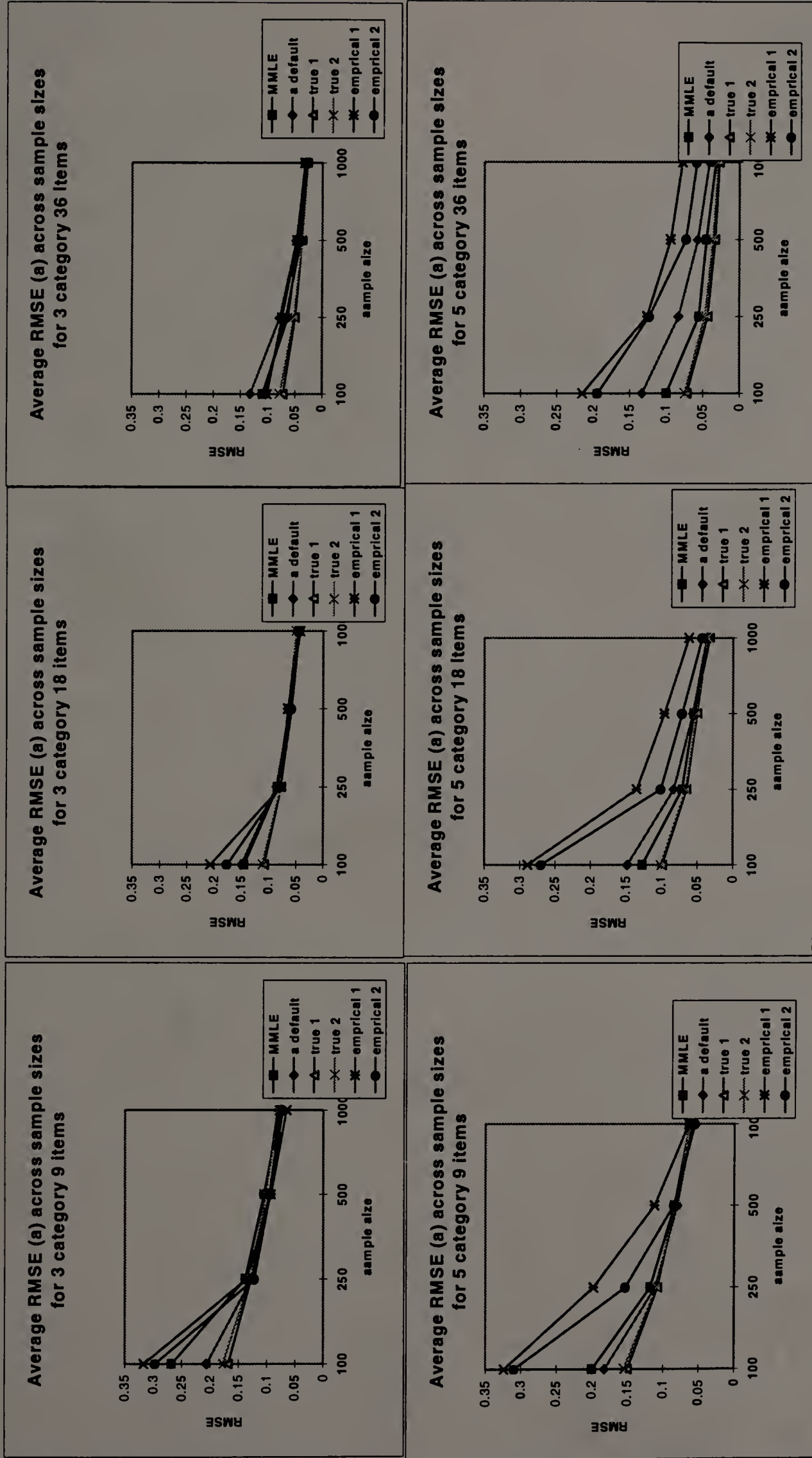


Figure 7. Average RMSE of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items

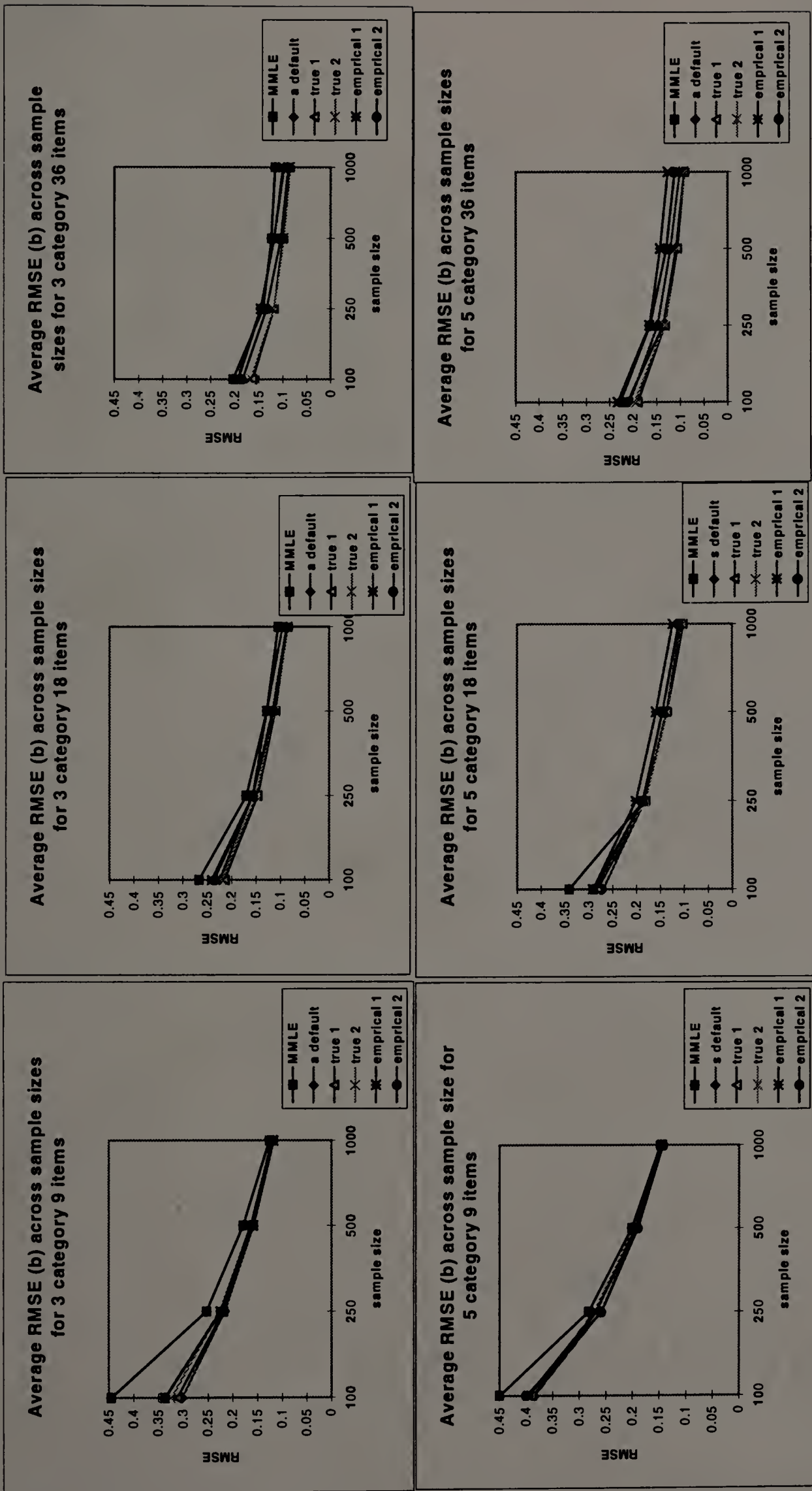


Figure 8. Average RMSE of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items

Figure 7 shows that, overall, priors based on the true distribution produced the smallest RMSE for estimates of the slope parameters. The default prior for the slope parameters, on the other hand, resulted in smaller RMSE than does MMLE in small data sets (9 and 18 items with 3 categories and 9 items with 5 categories), but yielded larger RMSE of estimates than MMLE in large data sets (36 items with 3 categories and 18 and 36 items with 5 categories). An explanation for this result is that despite the fact that the default priors did not match the distribution of the parameters, they were able to improve on MML in small data sets. With large data sets, the data overwhelmed the default priors while the MML procedure produced reasonably good estimates. Poor results were obtained with empirical priors. Empirical priors resulted in the largest RMSE for estimates of the slope parameters. In particular, empirical priors yielded larger RMSE for estimates of the slope parameters in 5-category items. This is in contrast to the results obtained with other prior distributions where smaller RMSE was obtained for the slope parameters of the five category items than with the three category items. It appears that the polyserial correlations, especially in the five category items showed a great deal of fluctuation from sample to sample, and did not reflect the true item parameter values. Consequently, empirical priors for the slope parameters based on the polyserial correlations produced poor results.

The effect of prior distributions on the estimation of the step difficulty parameters is clearer than that on the slope parameter. Figure 8 shows that using priors resulted in small RMSE for estimates of the step difficulty parameters, especially with small sample size. The MMLE procedure yielded larger RMSE for estimates of the category parameters than all Bayesian procedures. Even the generally poor performing empirical

priors reduced the RMSE of the step difficulty parameter estimates. These results show that MML procedure had more problems in the estimation of step difficulty parameters than in the estimation of slope parameters. Even poor specification priors on the slope parameters improved the estimation of step difficulty parameters.

5.3.2 Variance and Bias

In addition to the accuracy of estimates, the variance and bias of the estimates are important quantities in evaluating the quality of parameter estimation. The source of the difference between the estimates and the true parameter values, that is the accuracy of estimates, can be partitioned into sampling error, variance, and systematic bias. The sampling error is, in reality, the square of the standard error of the estimate obtained empirically. If an estimator shows great variation over repeated samples, i.e., has a large standard error, then the parameter will be estimated with less accuracy.

The average variance over 100 replications across all conditions for 3 and 5 category items is reported in Table 14 and 15. The results of the ANOVA of the average variance of the estimates of the slope and the mean of category parameters are reported in Table 16.

Table 16 shows that the number of categories in each item, test length, and sample size had an impact on the variance of estimates of the slope and the step difficulty parameters, but the prior distributions affected only the variance of the estimates of the slope parameters.

Table 14

Average variance of estimates of slope parameters across different priors
for 3 and 5 category items

			MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default
3 category	9 items	100	0.266	0.187	0.135	0.165	0.172	0.182	*****	*****
		250	0.13	0.119	0.103	0.113	0.098	0.118	0.119	*****
		500	0.095	0.089	0.082	0.087	0.079	0.088	0.089	0.094
		1000	0.063	0.06	0.058	0.06	0.057	0.06	0.06	0.062
	18 items	100	0.145	0.121	0.084	0.104	0.111	0.116	0.121	*****
		250	0.084	0.077	0.065	0.072	0.067	0.075	0.077	*****
		500	0.063	0.057	0.053	0.055	0.054	0.057	0.057	0.062
		1000	0.045	0.041	0.039	0.04	0.039	0.041	0.041	0.041
	36 items	100	0.105	0.092	0.061	0.074	0.075	0.088	0.092	*****
		250	0.055	0.057	0.046	0.051	0.054	0.055	0.057	*****
		500	0.045	0.039	0.034	0.036	0.038	0.038	0.039	0.038
		1000	0.032	0.027	0.025	0.026	0.027	0.027	0.027	0.027
5 category	9 items	100	0.197	0.164	0.118	0.145	0.166	0.169	0.162	*****
		250	0.114	0.105	0.095	0.1	0.104	0.108	0.107	0.114
		500	0.077	0.077	0.071	0.077	0.075	0.077	0.077	0.077
		1000	0.055	0.055	0.055	0.055	0.051	0.052	0.052	0.055
	18 items	100	0.122	0.114	0.078	0.095	0.109	0.113	0.111	*****
		250	0.071	0.071	0.055	0.063	0.067	0.069	0.069	0.071
		500	0.055	0.055	0.045	0.045	0.052	0.052	0.051	0.055
		1000	0.032	0.032	0.032	0.032	0.036	0.036	0.036	0.032
	36 items	100	0.089	0.089	0.060	0.073	0.084	0.09	0.093	*****
		250	0.045	0.051	0.045	0.045	0.053	0.053	0.051	0.051
		500	0.032	0.035	0.032	0.033	0.036	0.036	0.035	0.035
		1000	0.032	0.025	0.032	0.024	0.026	0.025	0.025	0.024

Table 15

Average variance of estimates of step difficulty parameters across different priors
for 3 and 5 category items

			MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default
3 category	9 items	100	0.431	0.278	0.298	0.315	0.261	0.272	*****	*****
		250	0.244	0.204	0.208	0.218	0.186	0.214	0.207	*****
		500	0.164	0.148	0.149	0.155	0.138	0.155	0.149	0.171
		1000	0.11	0.105	0.106	0.108	0.101	0.107	0.105	0.111
	18 items	100	0.257	0.181	0.202	0.214	0.193	0.207	0.202	*****
		250	0.155	0.128	0.134	0.141	0.134	0.146	0.129	*****
		500	0.102	0.095	0.097	0.1	0.092	0.099	0.095	0.116
		1000	0.074	0.069	0.07	0.071	0.068	0.075	0.069	0.073
	36 items	100	0.183	0.122	0.145	0.152	0.135	0.149	0.122	*****
		250	0.102	0.088	0.096	0.1	0.09	0.099	0.092	*****
		500	0.071	0.066	0.07	0.072	0.064	0.073	0.069	0.084
		1000	0.05	0.049	0.05	0.051	0.046	0.051	0.049	0.052
5 category	9 items	100	0.449	0.375	0.386	0.405	0.324	0.351	0.428	*****
		250	0.281	0.258	0.267	0.272	0.226	0.242	0.258	*****
		500	0.187	0.180	0.182	0.184	0.165	0.176	0.180	0.188
		1000	0.131	0.130	0.130	0.130	0.122	0.128	0.129	0.131
	18 items	100	0.223	0.240	0.248	0.252	0.224	0.23	0.241	*****
		250	0.179	0.162	0.174	0.175	0.15	0.161	0.163	0.182
		500	0.125	0.118	0.123	0.123	0.109	0.115	0.118	0.126
		1000	0.089	0.085	0.088	0.088	0.081	0.084	0.086	0.09
	36 items	100	0.205	0.201	0.212	0.219	0.142	0.205	0.199	*****
		250	0.122	0.106	0.119	0.121	0.1	0.101	0.107	0.127
		500	0.088	0.081	0.087	0.087	0.075	0.079	0.081	0.088
		1000	0.058	0.057	0.059	0.059	0.052	0.055	0.057	0.059

Table 16

Results of ANOVA for variance

Source	df	Estimates	F value	P value
Number of categories	1	Slope	12.24	.001
	1	Step difficulty parameters	6.84	.010
Test length	2	Slope	235.78	.000
	2	Step difficulty parameters	56.51	.000
Sample size	3	Slope	274.80	.000
	3	Step difficulty parameters	55.26	.000
Prior distributions	6	Slope	6.29	.000
	6	Step difficulty parameters	1.30	.261

Figures 9 and 10 provide summaries of the average variance of item parameter estimates for 3 and 5 category items across sample sizes (100, 250, 500 and 1000) and the effects of different priors (no prior, default 2, empirical prior 1 and 2, and true distribution based prior 1 and 2) at each test length. Both figures show that using any prior (including empirical priors) resulted in smaller variance of parameter estimates than using no prior. The variance of item parameter estimates decreased as the number of items, examinees, and categories in each item are increased.

If an estimator is unbiased, the mean of the estimates will converge to the true value as the number of replications approaches infinity. Consequently, the difference between the estimate and the true parameter value, MSD, is attributable to sampling error, or variance of the estimates. The average bias over 100 replications across all conditions for 3 and 5 category items is reported in Table 17 and 18.

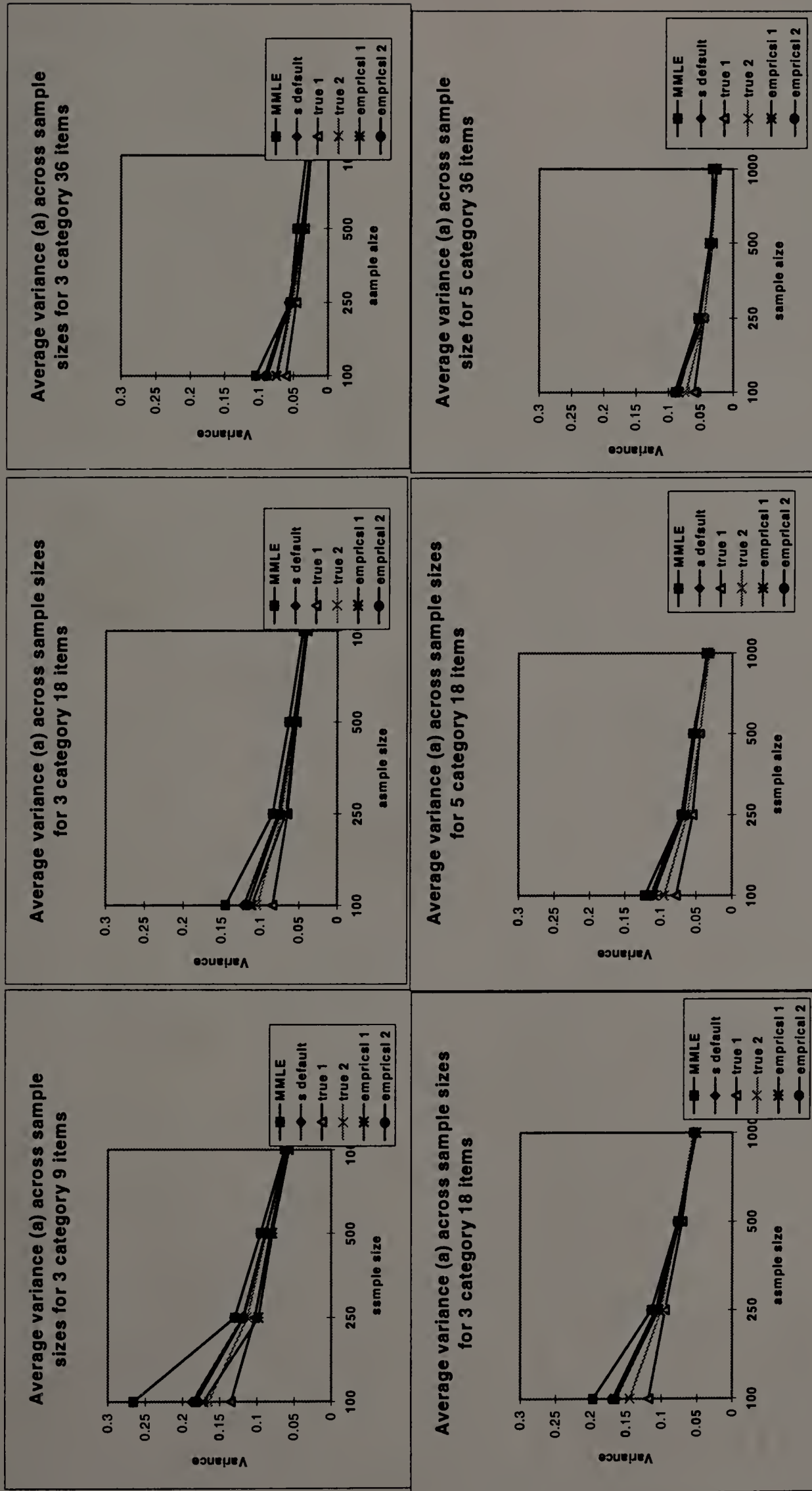


Figure 9. Average variance of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items

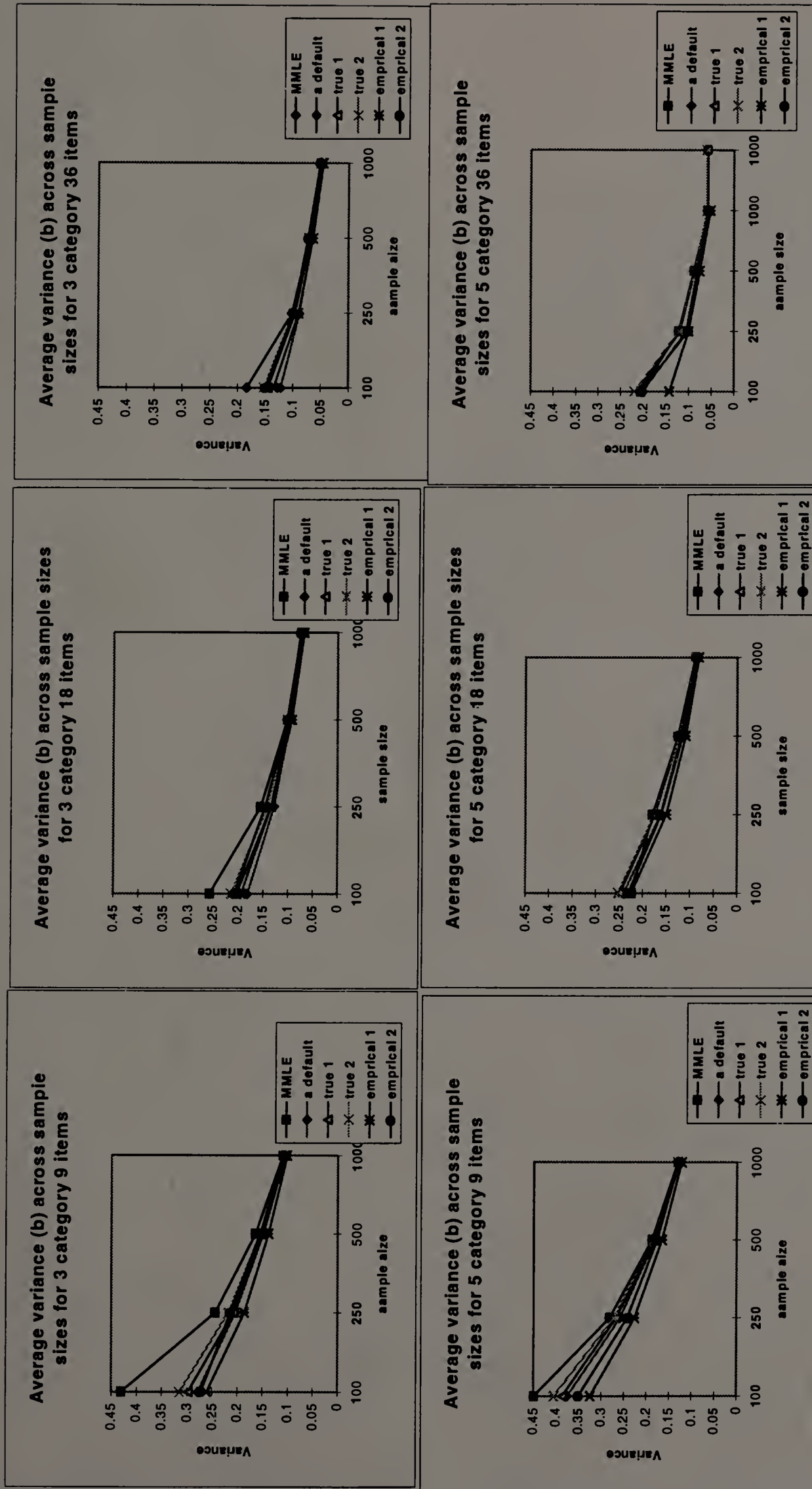


Figure 10. Average variance of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items

Table 17

Average bias of estimates of slope parameters across different priors
for 3 and 5 category items

			MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default
3 category	9 items	100	0.032	0.087	0.099	0.067	0.266	0.235	*****	*****
		250	0.045	0.043	0.075	0.061	0.094	0.033	0.043	*****
		500	0.045	0.037	0.06	0.054	0.049	0.032	0.037	0.047
		1000	0.045	0.044	0.057	0.054	0.031	0.04	0.044	0.051
	18 items	100	0.032	0.089	0.067	0.041	0.175	0.133	0.088	*****
		250	0	0.038	0.042	0.029	0.039	0.019	0.037	*****
		500	0	0.02	0.036	0.031	0.037	0.012	0.02	0.026
		1000	0	0.019	0.03	0.028	0.014	0.016	0.019	0.026
	36 items	100	0.045	0.095	0.038	0.022	0.067	0.063	0.094	*****
		250	0.032	0.053	0.019	0.011	0.05	0.038	0.051	*****
		500	0	0.028	0.013	0.008	0.027	0.013	0.027	0.007
		1000	0	0.015	0.009	0.007	0.018	0.008	0.015	0.005
5 category	9 items	100	0	0.077	0.095	0.055	0.278	0.26	0.078	*****
		250	0.032	0.032	0.055	0.045	0.169	0.109	0.030	0.032
		500	0.032	0	0.045	0.032	0.083	0.036	0.022	0.032
		1000	0.032	0.032	0.032	0.032	0.038	0.017	0.027	0.032
	18 items	100	0.031	0.095	0.055	0.032	0.268	0.245	0.089	*****
		250	0	0.045	0.032	0	0.118	0.073	0.045	0
		500	0	0	0.000	0	0.081	0.048	0.022	0
		1000	0	0	0.000	0	0.051	0.024	0.010	0
	36 items	100	0.045	0.099	0.021	0.019	0.197	0.171	0.101	*****
		250	0	0.065	0.000	0.013	0.115	0.111	0.065	0.018
		500	0	0.045	0.000	0.014	0.087	0.064	0.045	0.019
		1000	0	0.033	0.000	0.016	0.073	0.052	0.033	0.019

Table 18

Average bias of estimates of step difficulty parameters across different priors
for 3 and 5 category items

			MMLE	a default	true 1	true 2	emprical 1	emprical 2	default(ab)	b default
3 category	9 items	100	0.109	0.117	0.092	0.076	0.211	0.188	*****	*****
		250	0.07	0.072	0.067	0.06	0.124	0.065	0.07	*****
		500	0.067	0.055	0.056	0.056	0.075	0.055	0.054	0.067
		1000	0.063	0.053	0.055	0.056	0.058	0.054	0.053	0.058
	18 items	100	0.074	0.133	0.071	0.057	0.144	0.113	0.128	*****
		250	0.074	0.091	0.061	0.053	0.083	0.057	0.087	*****
		500	0.074	0.068	0.054	0.05	0.089	0.059	0.065	0.039
		1000	0.077	0.054	0.048	0.046	0.068	0.05	0.054	0.043
	36 items	100	0.092	0.152	0.074	0.066	0.136	0.109	0.151	*****
		250	0.097	0.118	0.07	0.065	0.115	0.091	0.108	*****
		500	0.102	0.097	0.069	0.066	0.098	0.075	0.092	0.031
		1000	0.107	0.084	0.069	0.067	0.089	0.073	0.083	0.061
5 category	9 items	100	0.081	0.104	0.102	0.096	0.227	0.215	0.101	*****
		250	0.07	0.071	0.073	0.071	0.155	0.108	0.070	0.07
		500	0.069	0.067	0.068	0.068	0.098	0.07	0.066	0.069
		1000	0.068	0.065	0.068	0.068	0.076	0.065	0.066	0.07
	18 items	100	0.249	0.131	0.338	0.335	0.184	0.168	0.31	*****
		250	0.052	0.087	0.061	0.106	0.137	0.098	0.086	0.041
		500	0.047	0.072	0.058	0.054	0.117	0.089	0.071	0.048
		1000	0.053	0.063	0.053	0.053	0.096	0.073	0.062	0.053
	36 items	100	0.059	0.35	0.392	0.383	0.183	0.161	0.352	*****
		250	0.059	0.103	0.059	0.058	0.132	0.13	0.103	0.049
		500	0.067	0.092	0.065	0.066	0.122	0.103	0.092	0.065
		1000	0.07	0.085	0.069	0.07	0.116	0.098	0.085	0.07

The results of the ANOVA for the average bias over all conditions for the estimates of the slope and the mean of category parameters are reported in Table 19.

Table 19 shows that test length and sample size influenced the bias in the estimates of the slope and step difficulty parameters. The number of categories in each item affected the bias in the estimates of the step difficulty parameters, but did not affect that in the estimates of the slope parameters. Prior distributions influenced the bias in the estimates of slope parameters, but did not influence that of step difficulty parameters.

Table 19
Results of ANOVA for Bias

Source	df	Estimates	F value	P value
Number of categories	1	Slope	2.65	.106
	1	Step difficulty parameters	18.16	.000
Test length	2	Slope	7.65	.001
	2	Step difficulty parameters	5.01	.008
Sample size	3	Slope	41.19	.000
	3	Step difficulty parameters	37.58	.000
Prior distributions	6	Slope	16.05	.000
	6	Step difficulty parameters	1.83	.097

Figures 11 and 12 provide summaries of the average bias of estimates of item parameters for 3 and 5 category items across sample sizes (100, 250, 500 and 1000) and different priors (no prior, default prior 2, empirical prior 1 and 2, and true distribution

based prior 1 and 2) for each test length. As can be expected, using priors provided larger bias than not using priors. The bias in the estimates of slope parameters from empirical priors was much larger than that of other priors; in particular, the bias in the estimates for 5-category items is much larger than that for the 3-category items.

The number of items and the type of prior seem to affect the bias in the estimates of the slope parameters. That is, true distribution-based priors resulted in large bias in the estimates of slope parameters in short tests (9 items) and default priors resulted in large bias in the estimates of slope parameters in longer tests (18 and 36 items). Differences in bias among priors decreased as sample sizes increased. However, the bias in the estimates of step difficulty parameters did not become zero even when the sample size became very large. Also, the bias in estimates of category parameters in small sample size for longer tests (5-category 18 and 36 item tests) became large with true distribution-based priors.

The pattern for bias in the estimates of the parameters is parallel to that observed for the RMSE. This can be explained in terms of the decomposition of MSD into variance and bias. Since the variance term was small, large MSD and hence RMSE was the result of bias in the estimates.

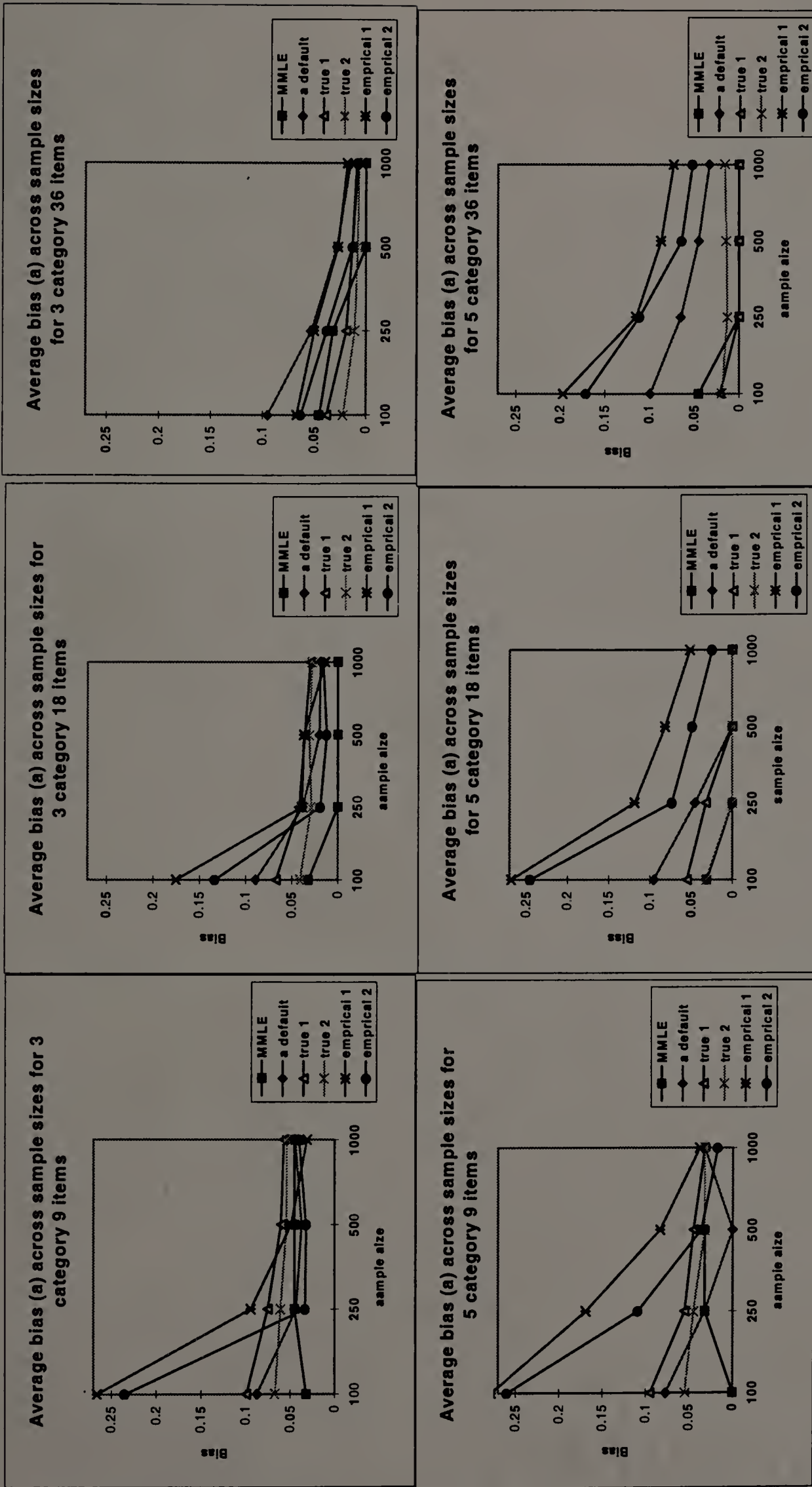


Figure 11. Average bias of estimates of slope parameters across sample sizes and different priors for 3 and 5 category items

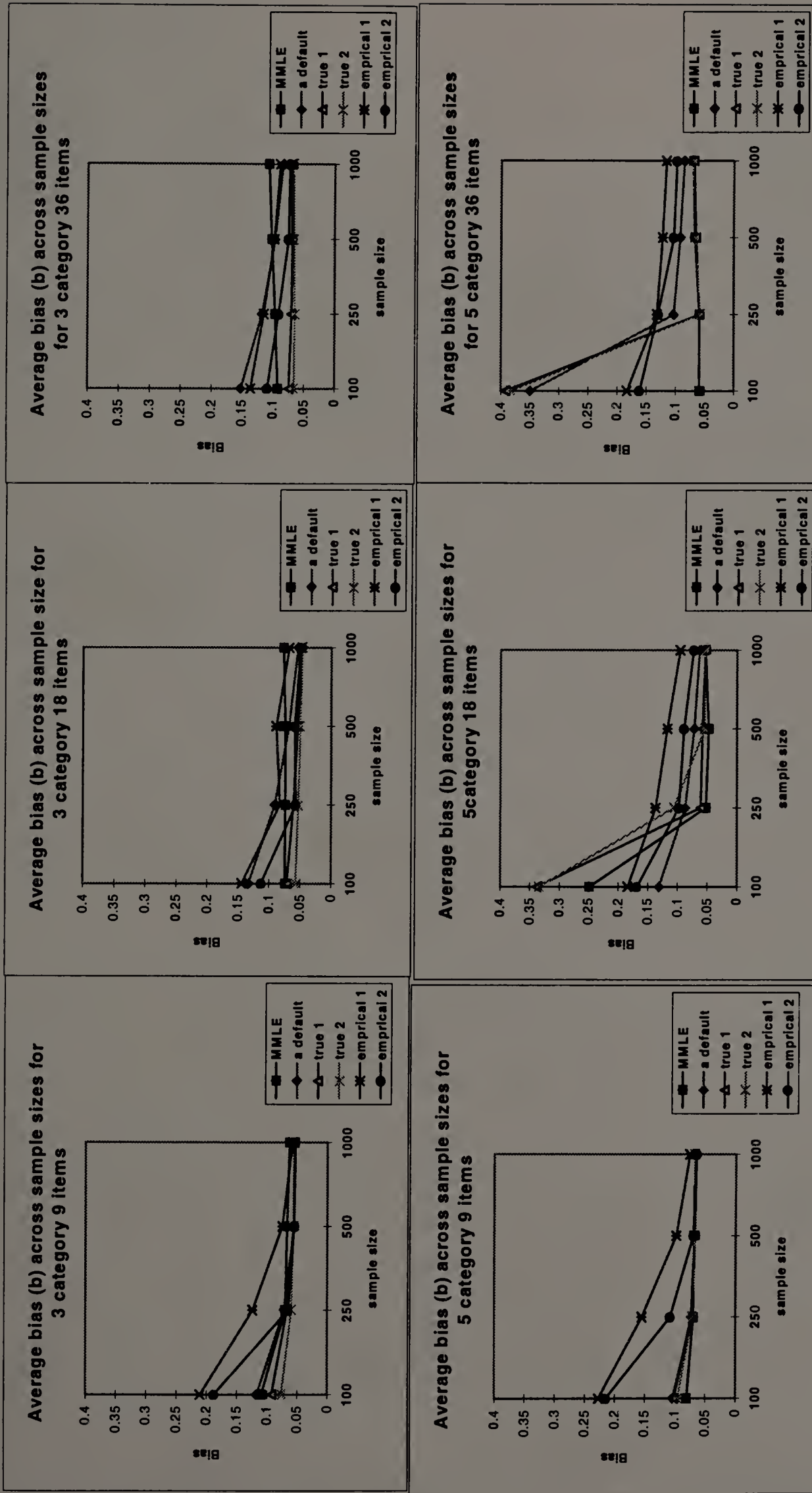


Figure 12. Average bias of estimates of step difficulty parameters across sample sizes and different priors for 3 and 5 category items

5.3.3 Item Level Analysis on the Accuracy of Estimation

To examine the effects of item parameter values and prior distributions on accuracy of estimates of item parameters, item level analysis was performed. Item level analysis involves examining the nine types of items, explained earlier in the context of data generation. The true distribution-based priors can be matched to middle levels of slope parameter items and default priors can be matched to high levels of slope parameter items.

The RMSE of estimates for each item type is reported only for 9- and 18- item tests, since the priors did not seem to have any effect on the RMSE of estimates for the 36-item test. In the Figures that follow, item types are represented on the x-axis. Each item type is characterized by two letters; the first letter (L=Low, M = Medium, or H=High) represents the level of slope parameter value; the second letter (L=Low, M=Medium, or H=High) represents level of the step difficulty parameter value.

The RMSE of estimates of item parameters (slope parameter followed by step difficulty parameter) over replications is shown in Figures 13, 14, and 15 for a 9-item test with three categories. Figures 16, 17, and 18 show the RMSE of estimates of item parameters over replications for a three-category 18-item test. Figures 19, 20, 21, 22 and 23 show RMSE of estimates of item parameters over replications for a 5- category 9- item test while Figures 24, 25, 26, 27, and 28 show the RMSE of estimates of item parameters for 5- category 18 item-test.

In general, the accuracy of estimation was affected by the prior distribution and the item type. True distribution-based priors produced smaller RMSE than default priors and MMLE for low, medium, and high level slope parameters; MMLE produced smaller

RMSE than default priors at low level slope parameters. However, default priors produced smaller RMSE than MMLE at medium and high level slope parameters values. Empirical priors produced the largest RMSE and hence the least accurate estimation of the slope parameters. Clearly, polyserial correlations are not good choice for specifying the mean of the prior distributions of the slope parameters. The effect of priors diminished as the sample size and test length increased, a result that is consistent with the fact that when large amounts of data swamp the priors. The item type also had an effect on the accuracy of estimation; slope parameters with low values were estimated more accurately than slope parameters with medium and high values. Slope parameters with high values were most poorly estimated.

While the effect of priors on the slope parameters was modest, specifying priors for the slope parameters had a positive effect on the estimation of step difficulty parameters. Figures 14, 15, 16, 18, 20, 21, 22, 23, 24, 26, 27, and 28 reveals that using priors for slope parameters reduced RMSE of estimates of step difficulty parameters. Even Empirical priors resulted in smaller RMSE of estimates of step difficulty parameters than MMLE. It is interesting to note that empirical priors for the slope parameters resulted in the largest RMSE of estimates of the slope parameters, but yielded smaller RMSE of estimates of step difficulty parameters than MMLE. It appears that any prior on the slope parameters increased the estimation accuracy of step difficulty parameters, even if it did not improve the estimation of the slope parameter!

The type of item had an effect on the accuracy of estimation of the step difficulty parameters. Extremely low or high values of step difficulty parameters were estimated poorly by MMLE; using priors improved the estimation accuracy of this type of items.

The level of the slope parameter values also had an effect on the estimation of step difficulty parameters; low level of slope parameters resulted in poor estimation of step difficulty parameters. This effect was particularly noticeable for sample sizes less than 250 with MMLE producing the largest RMSE for the estimates of step difficulty parameters.

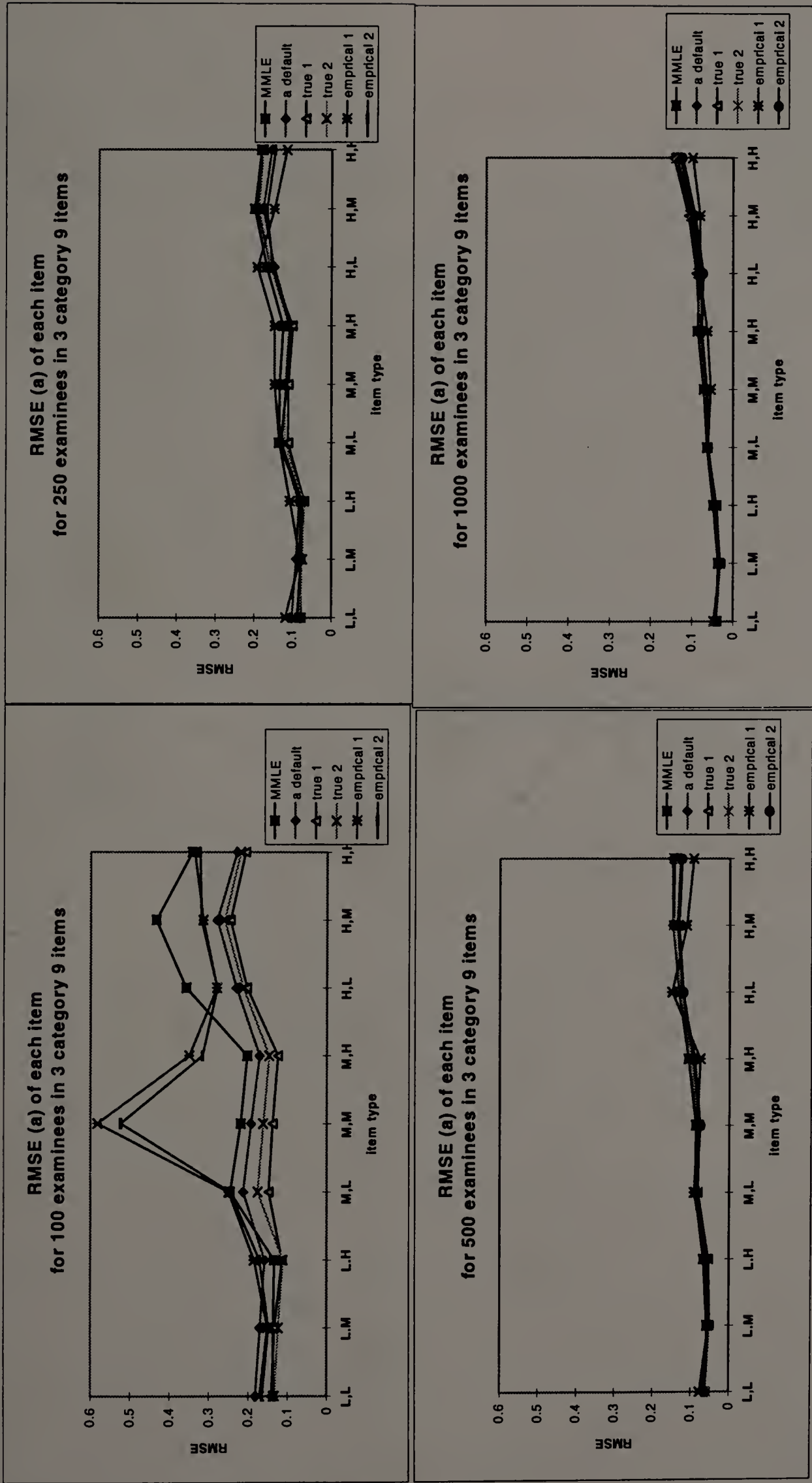


Figure 13. RMSE of estimates of slope parameters for each item in 3 category 9 items

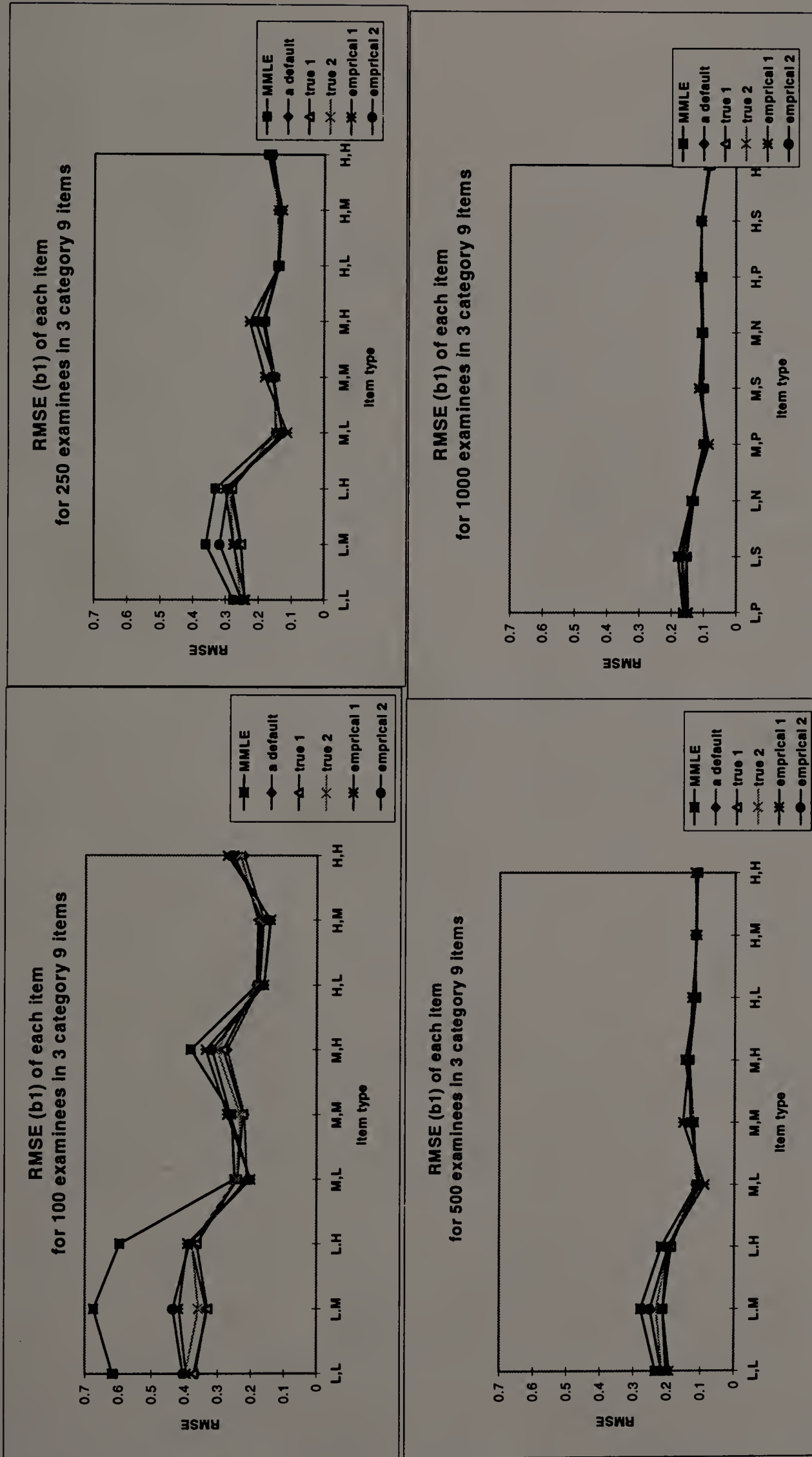


Figure 14. RMSE of estimates of the first step difficulty parameters for each item in 3 category 9 items

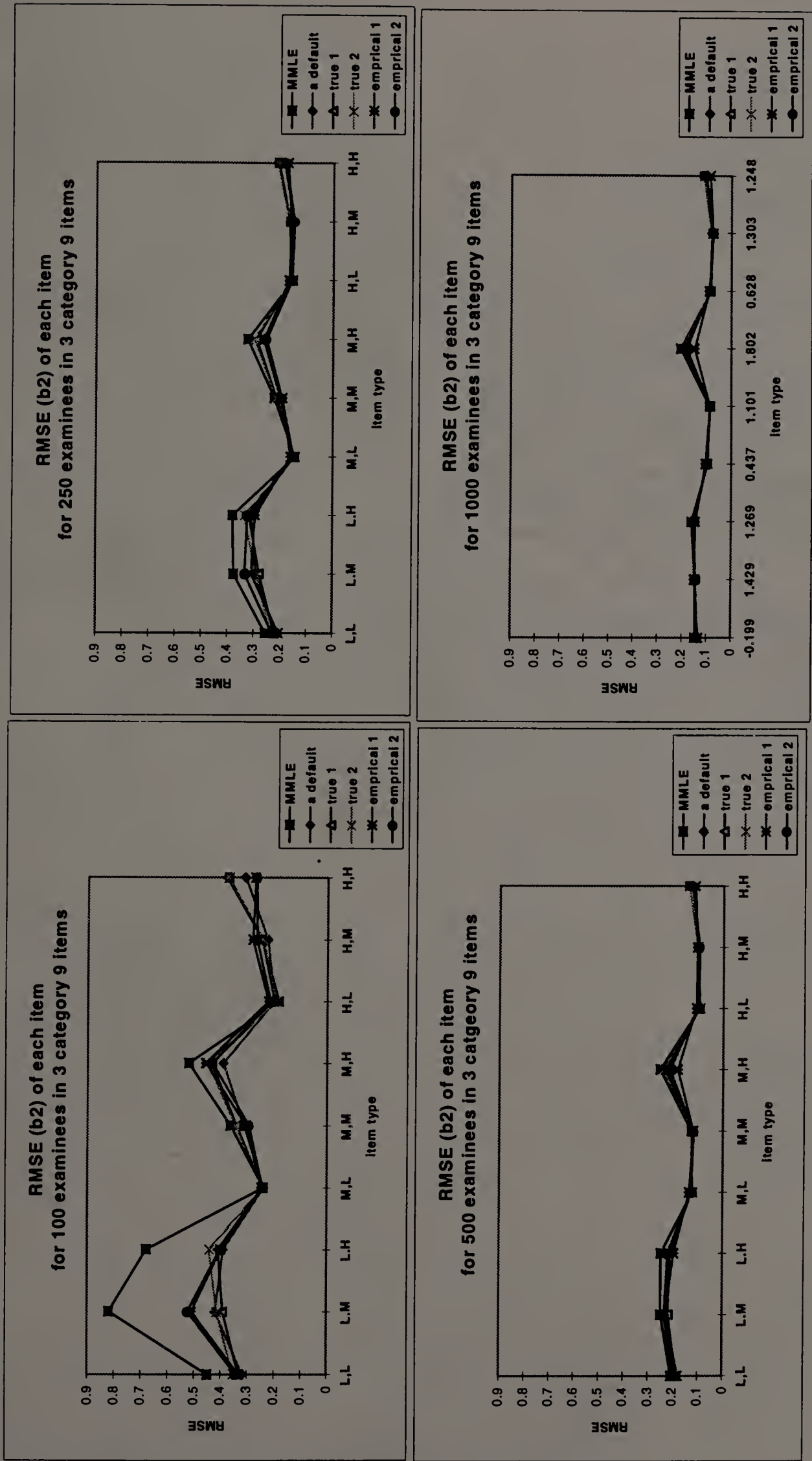


Figure 15. RMSE of estimates of the second step difficulty parameters for each item in 3 category 9 items

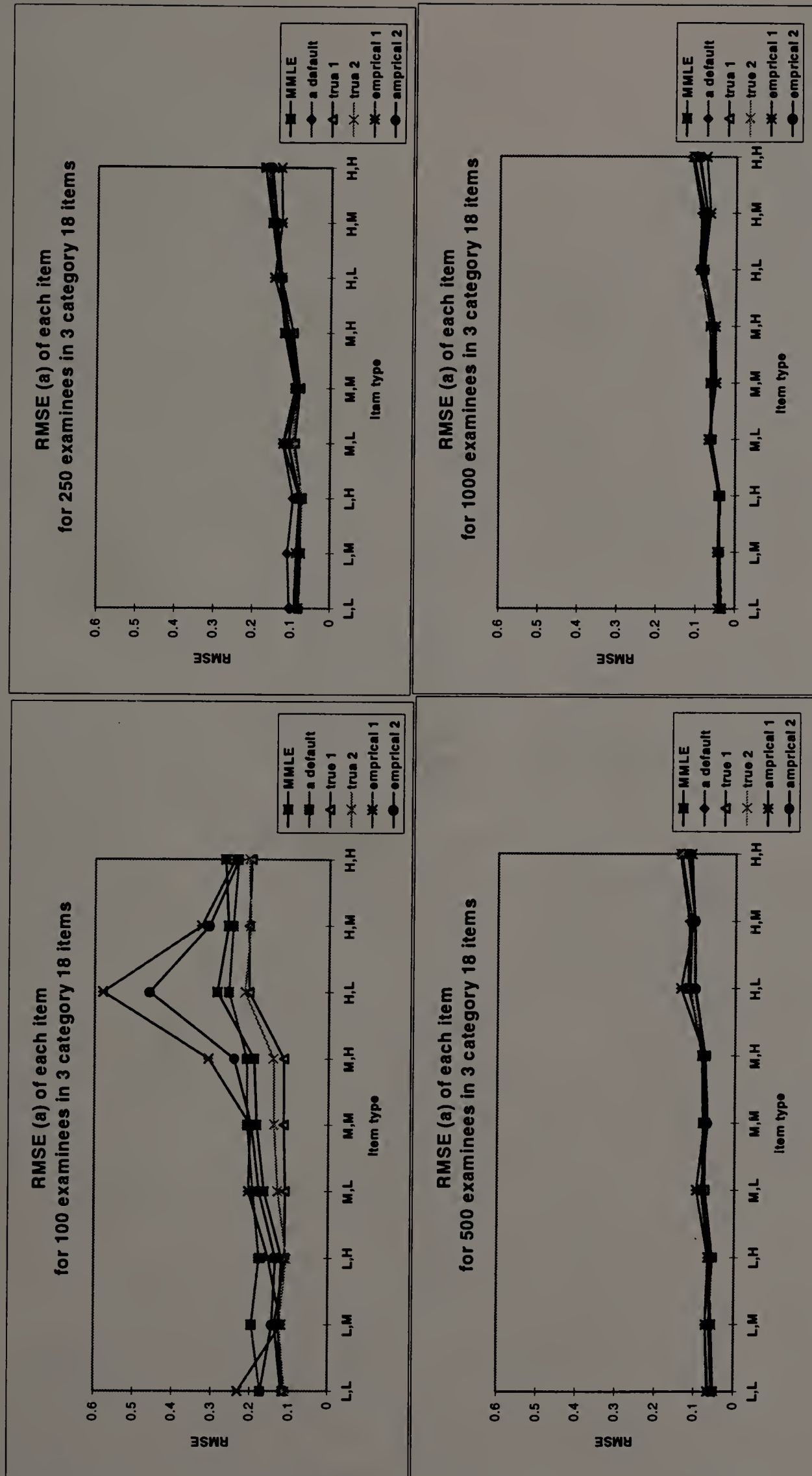


Figure 16. RMSE of estimates of slope parameters for each item in 3 category 18 items

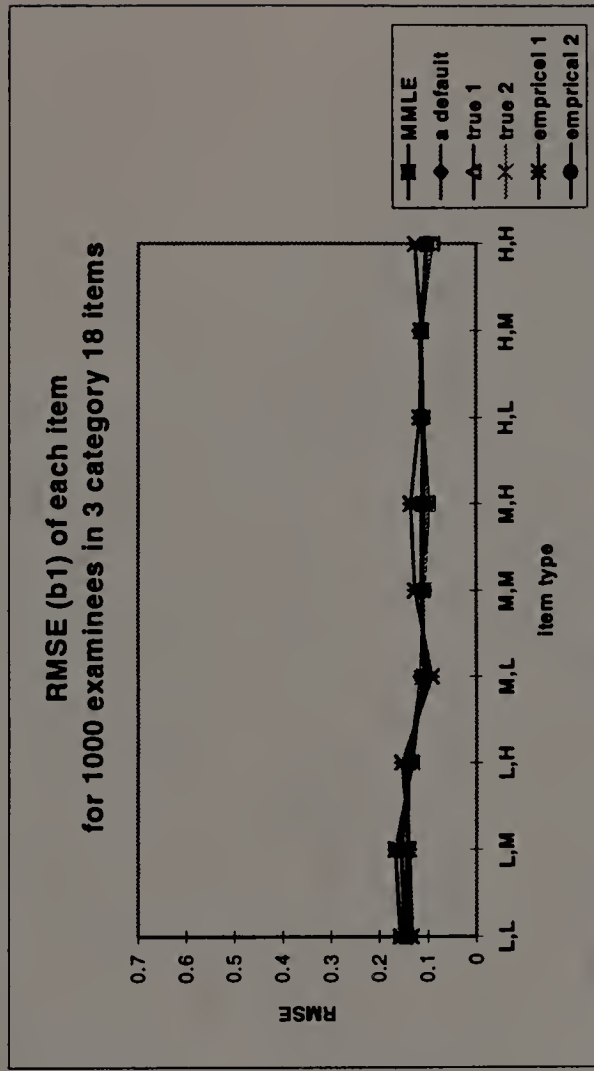
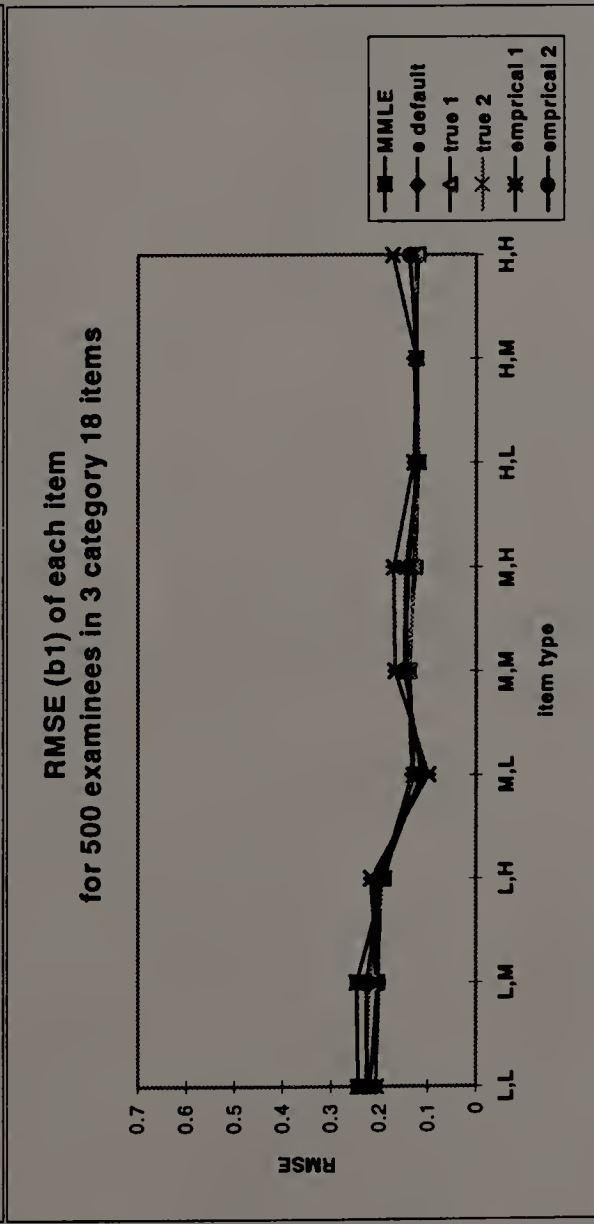
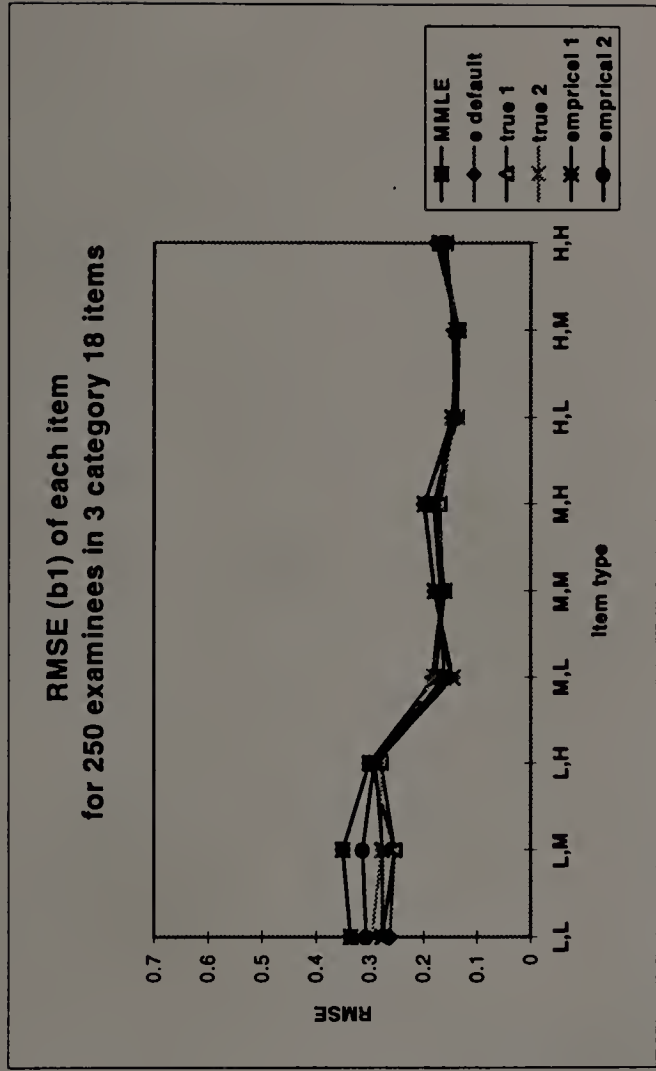
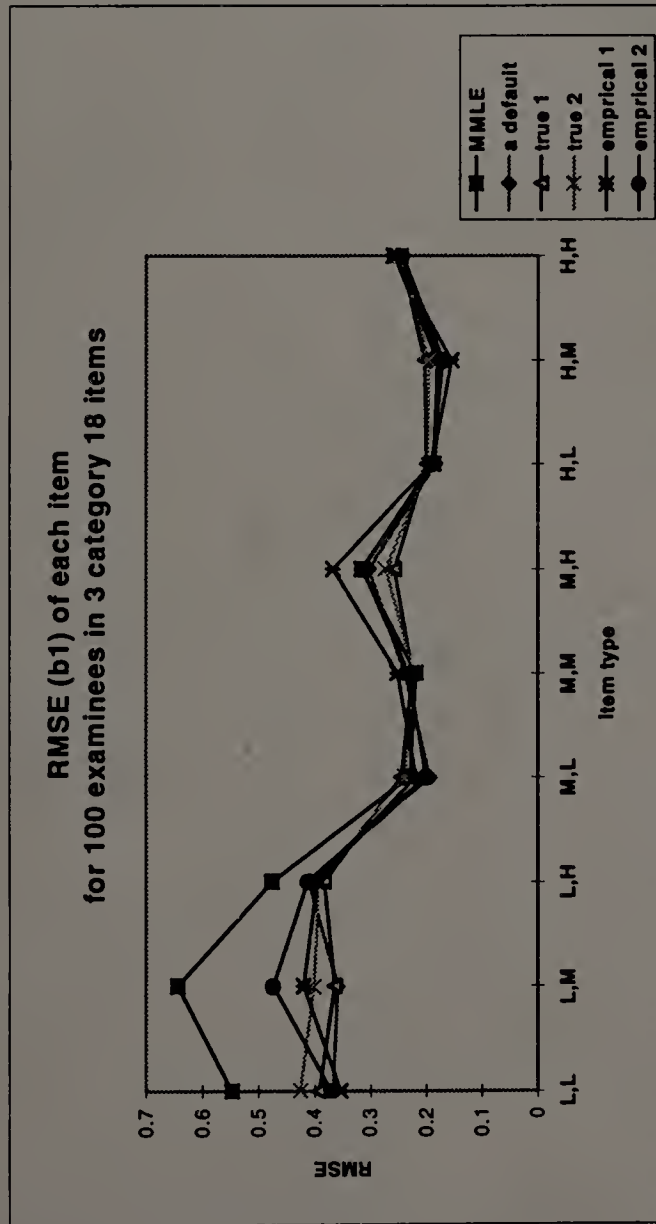


Figure 17. RMSE of estimates of the first step difficulty parameters for each item in 3 category 18 items

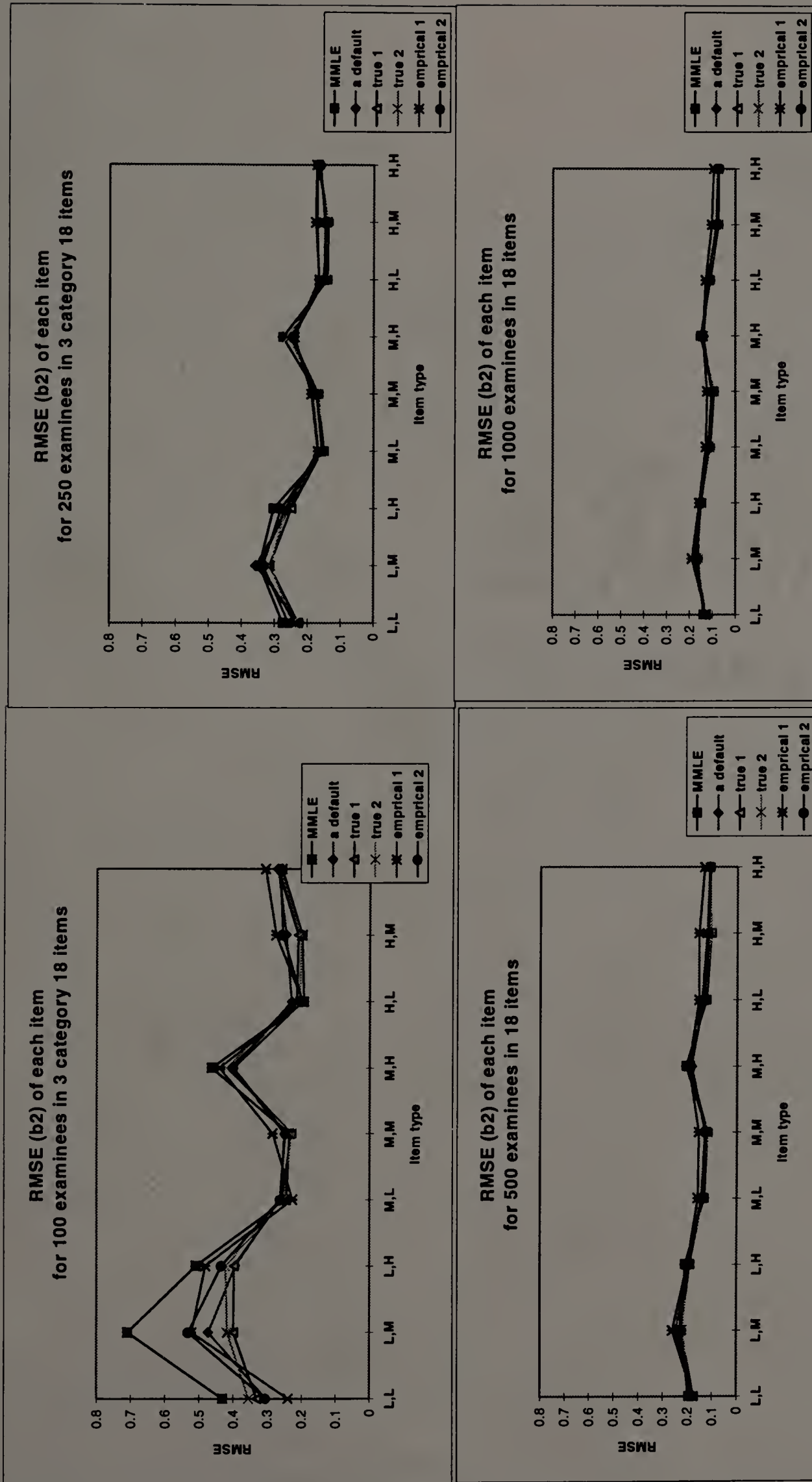


Figure 18. RMSE of estimates of the second step difficulty parameters for each item in 3 category 18 items

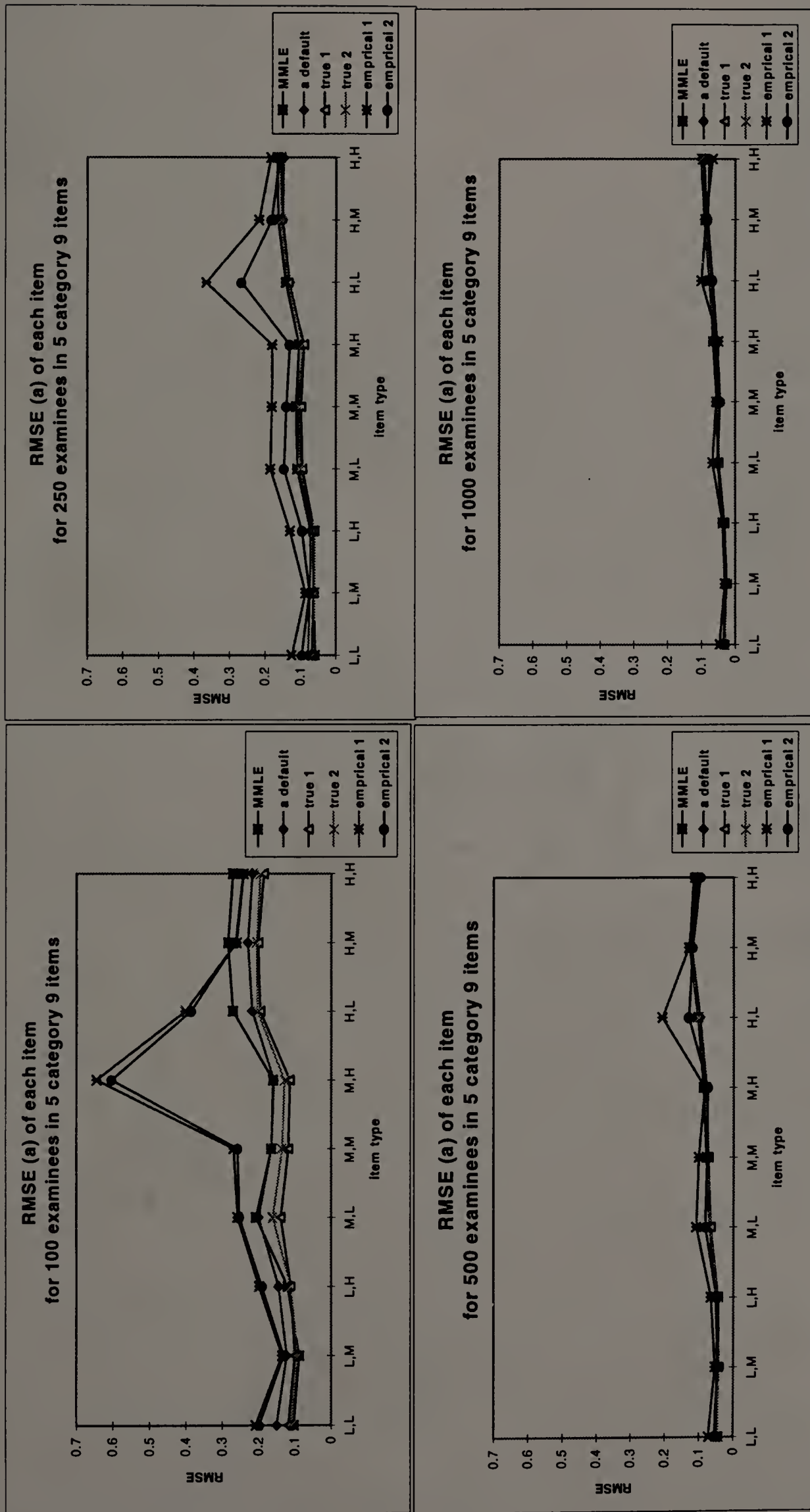


Figure 19. RMSE of estimates of slope parameters for each item in 5 category 9 items

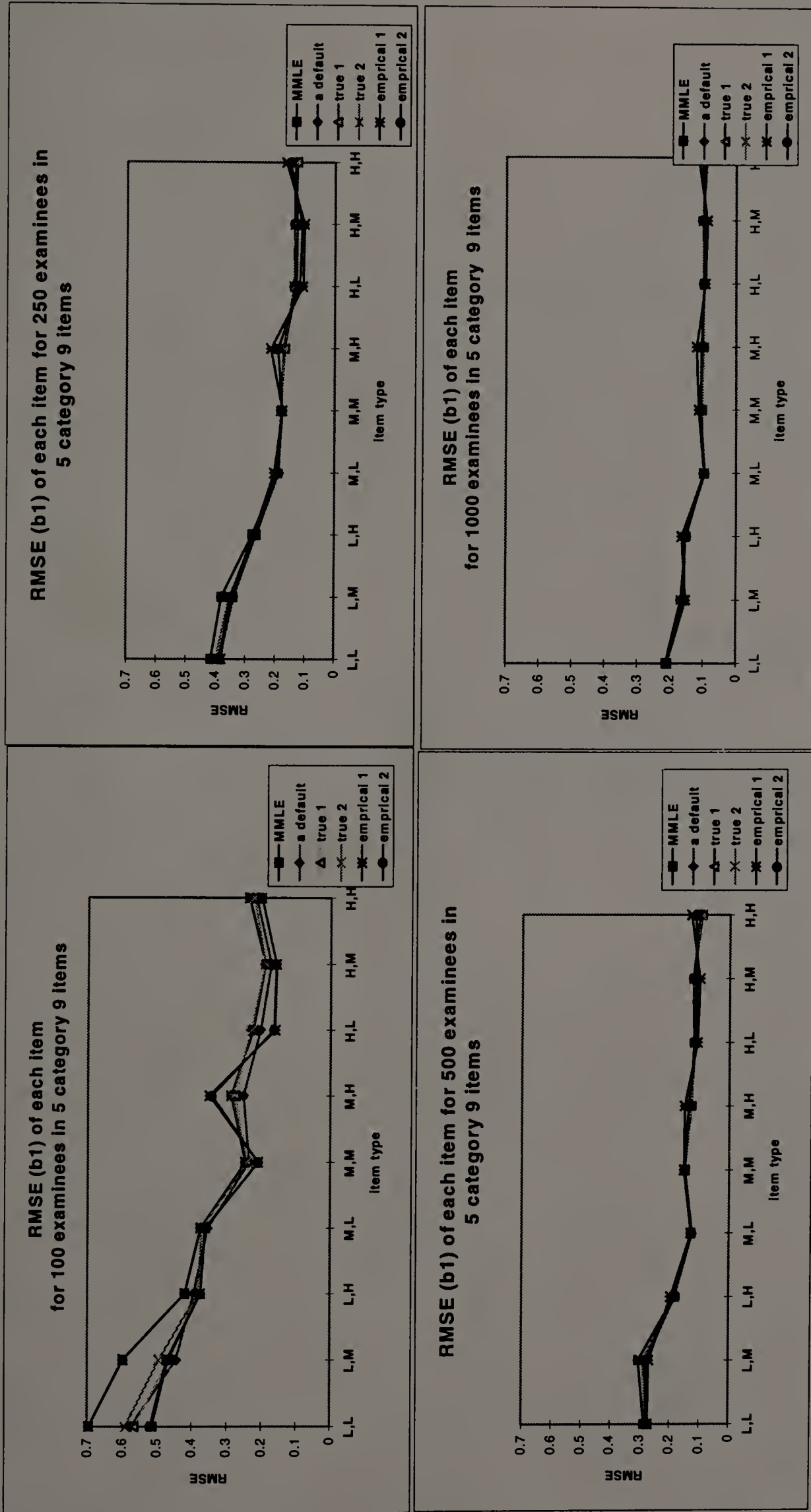


Figure 20. RMSE of estimates of the first step difficulty parameters for each item in 5 category 9 items

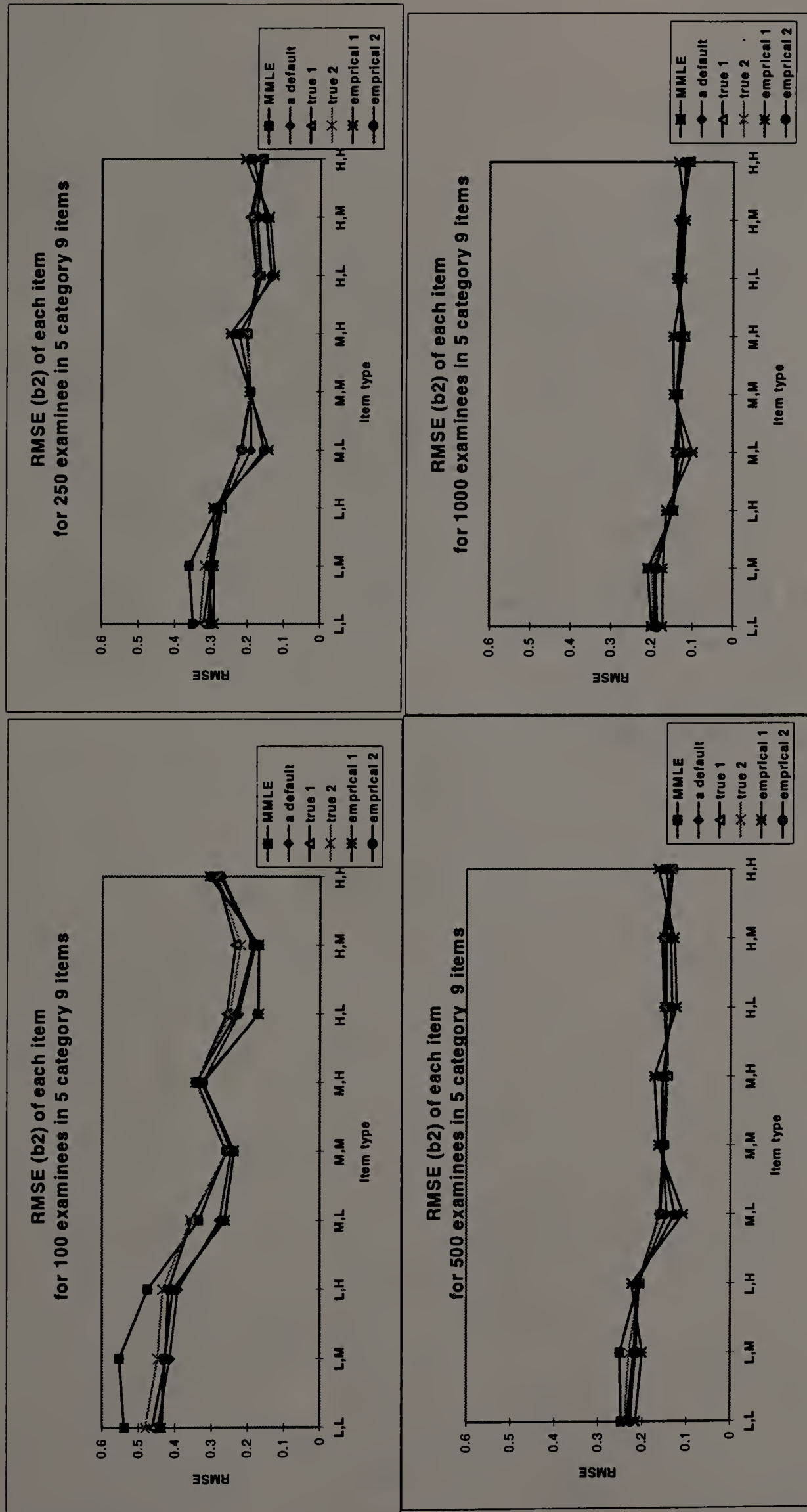


Figure 21. RMSE of estimates of the second step difficulty parameters for each item in 5 category 9 items

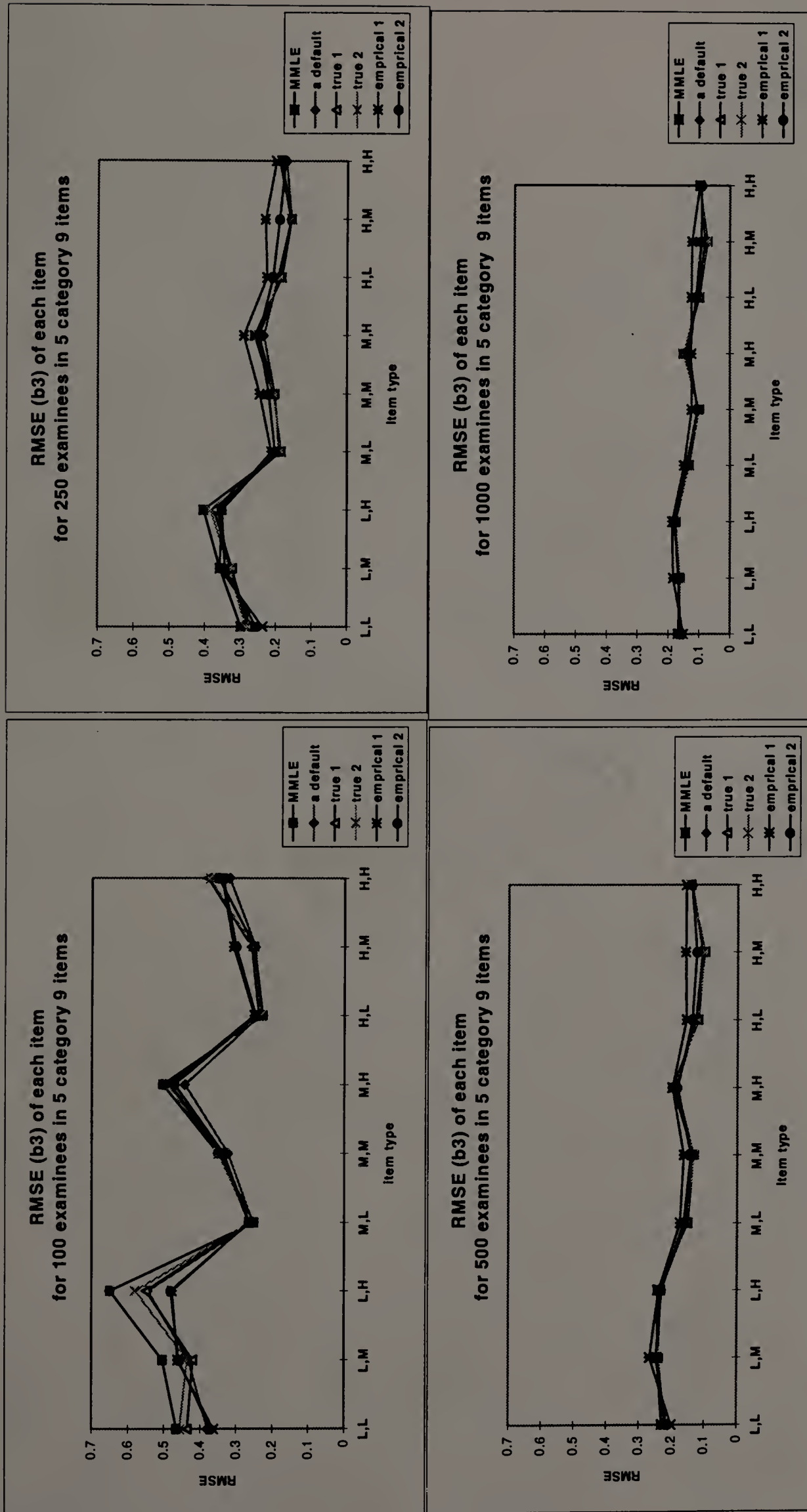


Figure 22. RMSE of estimates of the third step difficulty parameters for each item in 5 category 9 items

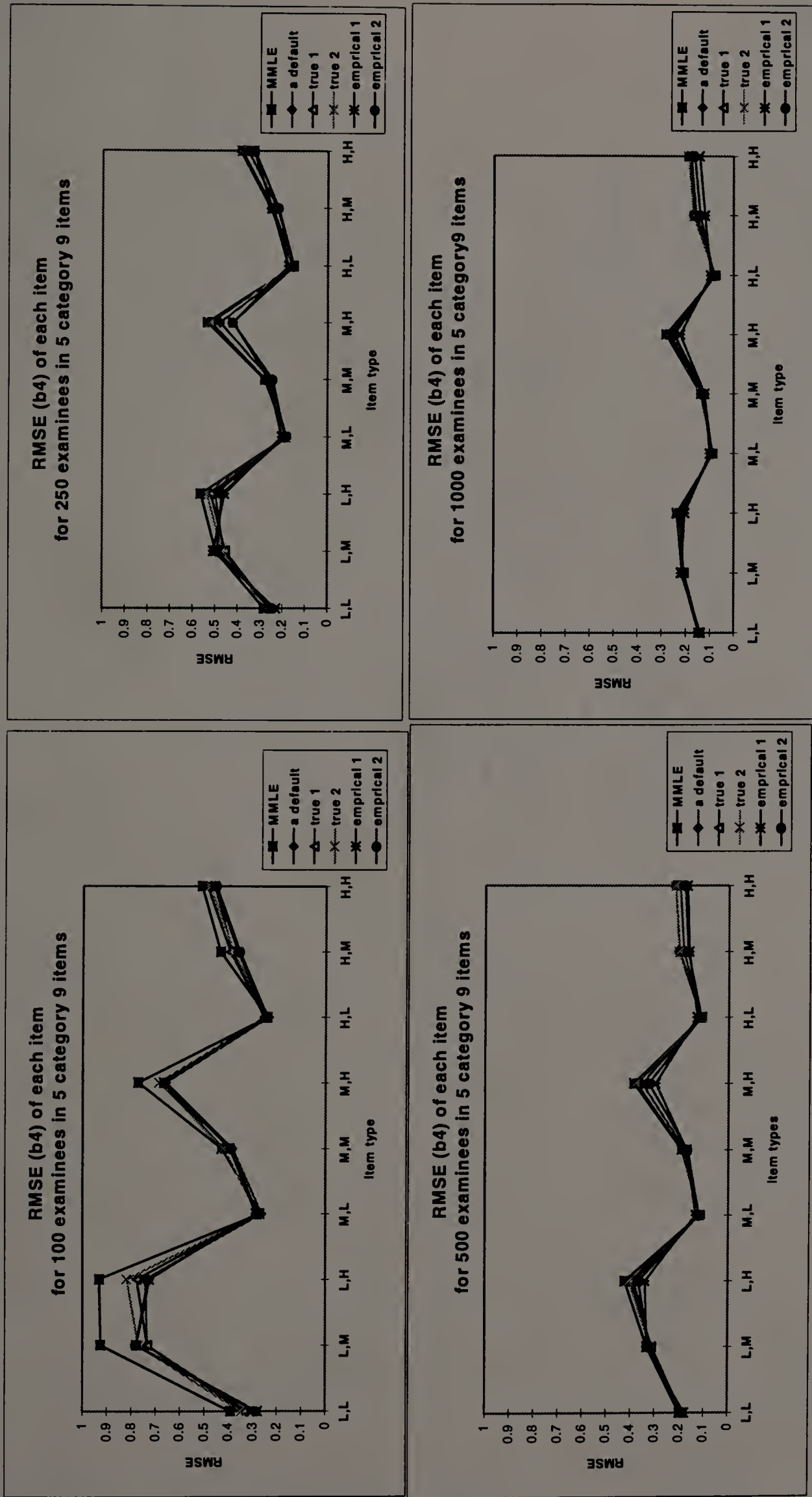


Figure 23. RMSE of estimates of the fourth step difficulty parameters for each item in 5 category 9 items

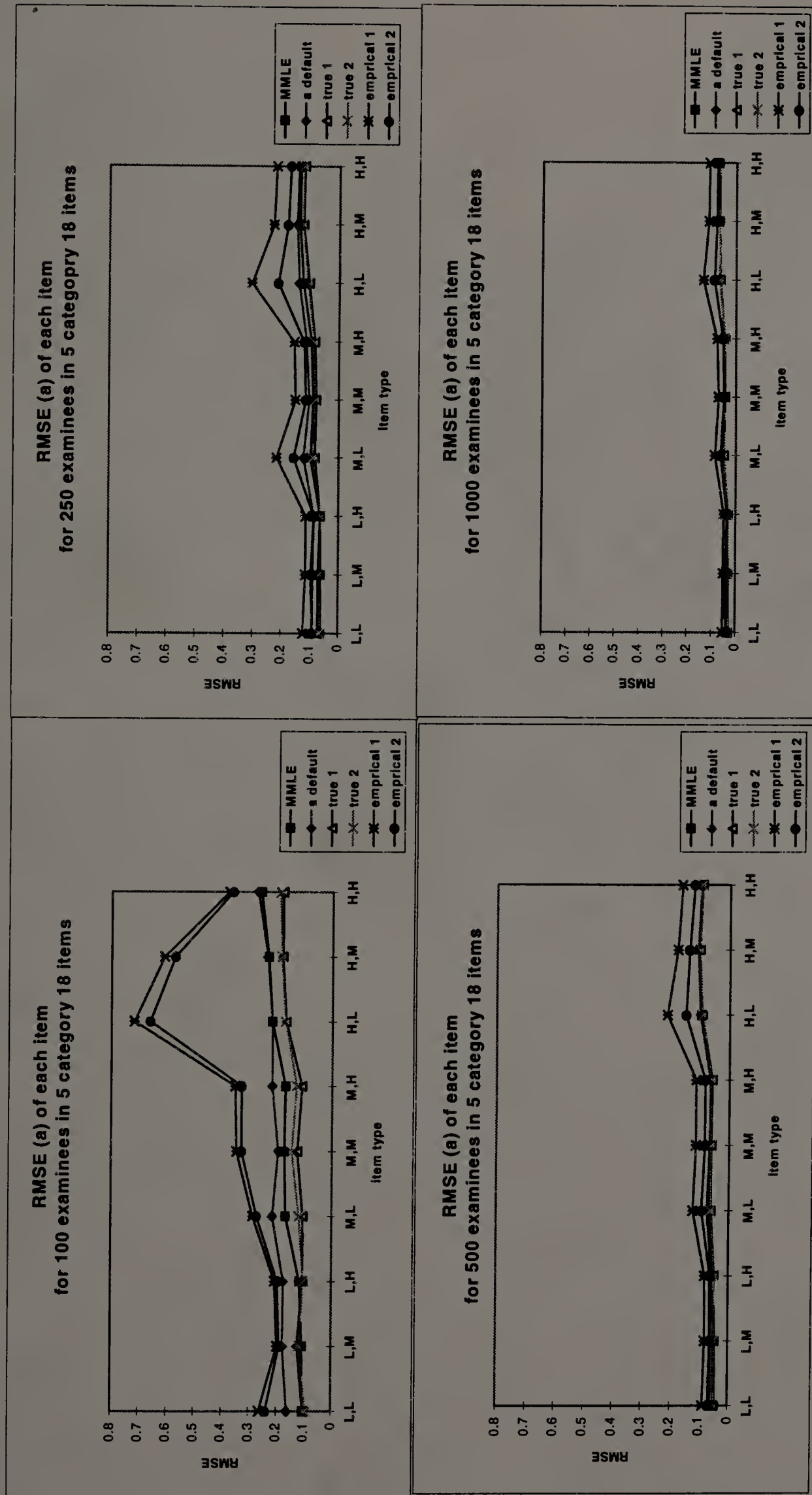


Figure 24. RMSE of estimates of slope parameters for each item in 5 category 18 items

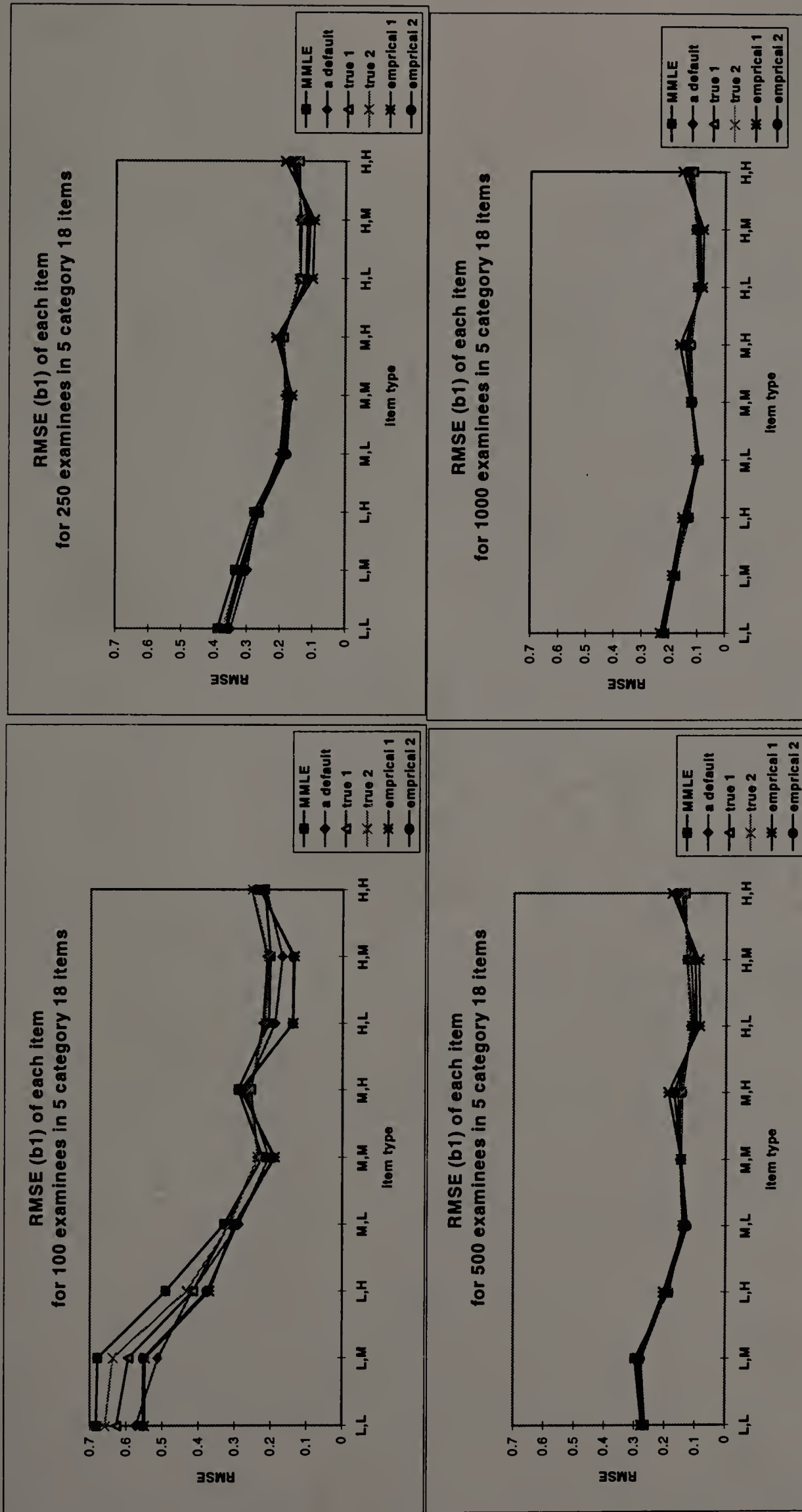


Figure 25. RMSE of estimates of the first step difficulty parameters for each item in 5 category 18 items

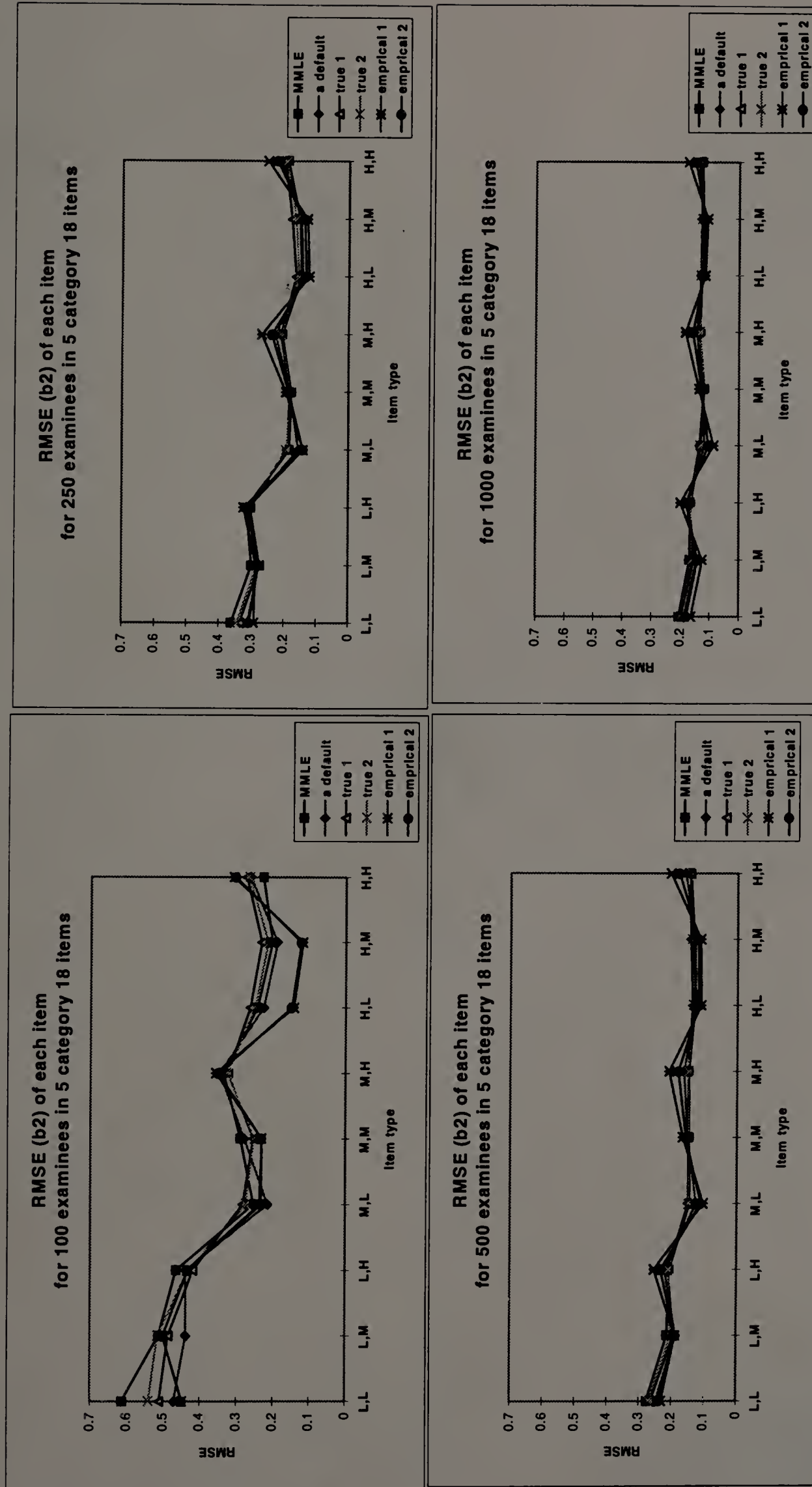


Figure 26. RMSE of estimates of the second step difficulty parameters for each item in 5 category 18 items

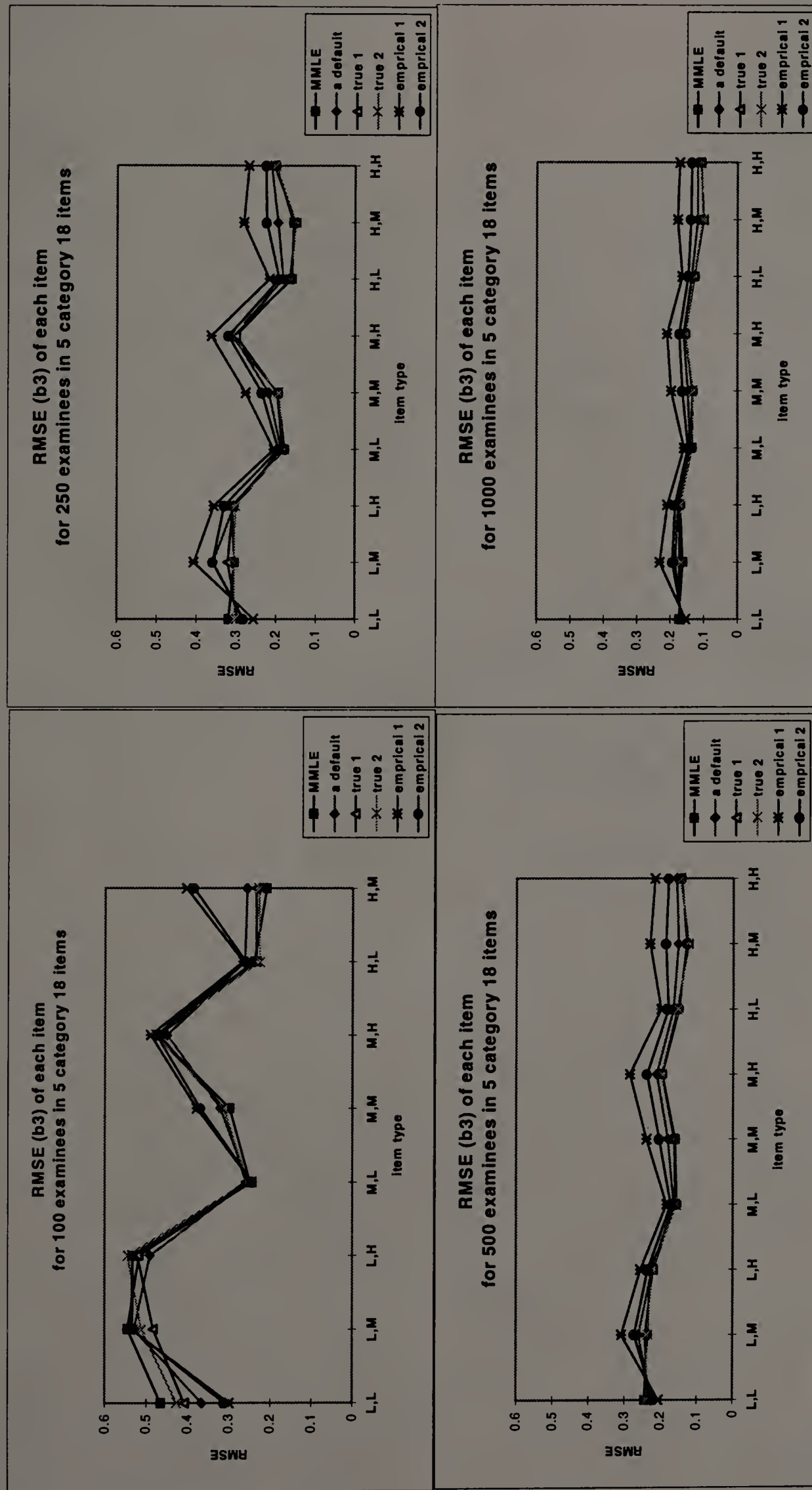


Figure 27. RMSE of estimates of the third step difficulty parameters for each item in 5 category 18 items

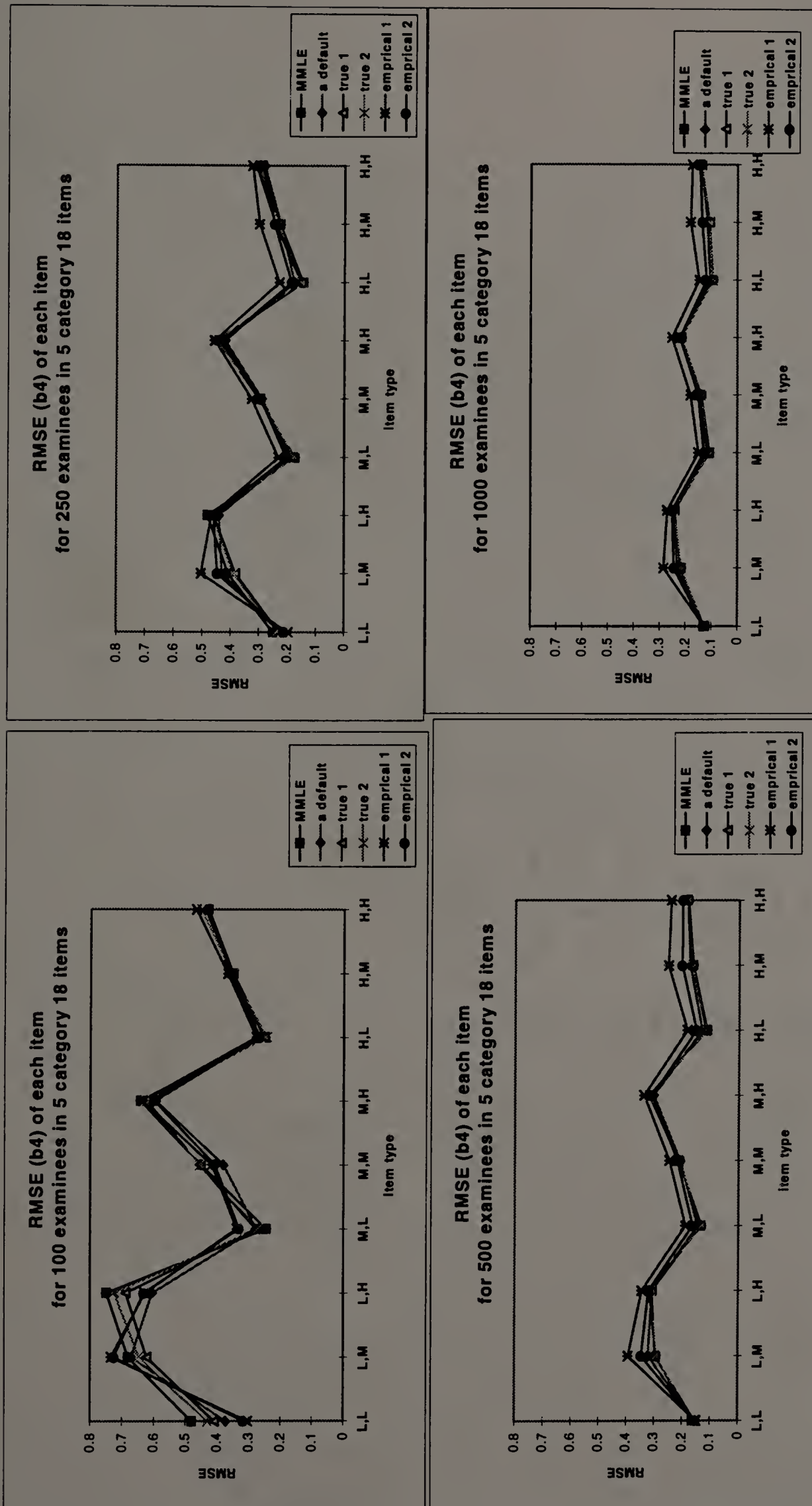


Figure 28. RMSE of estimates of the fourth step difficulty parameters for each item in 5 category 18 items

CHAPTER 6

SUMMARY AND CONCLUSIONS

In the previous chapter, the results of study I and II were presented in detail. This chapter contains: (a) a summary of important findings and conclusions from study I and II; (b) significance of the results; and (c) delimitations of the study and directions for further research.

6.1 Summary and Conclusions for Study I

In study I, the effects of sample size, test length, the number of categories in each item, and ability distribution on the MML estimates in the GPCM were investigated. The results showed that the ability distribution did not have an effect on the accuracy and the sampling fluctuations of the estimation of the parameters but had an impact on the bias of estimates of item parameters. Uniform ability distribution yielded the largest bias for both item parameters. Negatively skewed distribution produced the smallest bias for the slope parameter.

Increasing number of categories from three to five had a positive effect on the estimation of the slope parameters. As the number of categories increased, the RMSE and the variance of estimates in the slope parameters decreased. However, the number of categories in each item did not seem to affect the bias of estimates in the slope parameters. On the other hand, increasing number of categories had a negative effect on the estimation of the step difficulty parameters. As the number of categories increased,

the RMSE, the variance and the bias of estimates in the step difficulty parameters increased.

An explanation of the above result is that for a fixed number of examinees, increasing the number of categories results in fewer examinees in each category and this in turn affects the accuracy of estimation of the category parameters. The slope parameter, being common across the categories, is not affected by the decrease in the number of examinees in each category. In fact, increasing the number of categories had a modest positive effect on the estimation of the slope parameters - the additional number of categories seems to provide more information for the estimation of the slope parameter. To examine this phenomenon further, data were generated for a dichotomous item response model with the same slope parameter and with the mean of the category parameters being the difficulty parameter. Results not reported here revealed that RMSE, variance, and bias in the slope parameter showed an increase when compared with the results for the three and five category items.

Results from study I showed that sample size and test length had a clear effect on the accuracy of estimation and sampling fluctuations of the estimates of parameters in the GPCM. As sample size and test length increased, the accuracy of estimates increased and the variance of estimates and the bias of estimates in the slope parameters decreased, but the bias of estimates in the step difficulty parameters remained constant. Even as sample size and test length increased, the bias of estimates in the step difficulty parameters increased.

The most noticeable decrease of RMSE and the variance of estimates of parameters occurred when sample size increased from 100 to 250. The improvement

beyond 250 was modest. The results of study I suggest that a minimum number of 250 examinees is required to obtain reasonably accurate parameter estimates of the GPCM with 3- and 5-category items with the computer program PARSCALE.

The results of the study I showed that, in general, the ability distribution did not affect the accuracy and variance of parameter estimates for the GPCM. This result, however, does not agree with the results of previous studies (Reise & Yu, 1991; De Ayala, 1995). Both these studies found that estimation was more accurate with samples drawn from uniform distributions than with samples from other distributions. These contradictory findings may be partly due to the fact that the item parameter values that were used in this dissertation were taken from a NAEP administration, and these items favored a negatively skewed distribution. The previous studies also used a different polytomous IRT model and a different computer program. This fact may have also contributed to the contradictory findings.

While the ability distribution did not influence the accuracy and variance of estimates, it did influence the bias of item parameter estimates. The bias of estimates was large with samples from a uniform distribution and small with samples from a negatively skewed distribution. This results may have been due to the fact that the true item parameter values used in the study had a negatively skewed distribution which matched the distribution of the true parameter values.

Overall, the results of the study I showed that MML estimators of the parameters of the GPCM, as obtained through the computer program, PARSCALE, performed well under various conditions. Even with a sample size as small as 250, reasonable parameter estimates of the GPCM can be obtained if there are some examinees in each category.

However, there was some bias in the estimates of the category parameters under all conditions. The average bias did not decrease when sample size and test length increased. Since the Mean squared error is the sum of sample variance and squared bias and the sampling variance decreased as sample size and test length increased, the bias was contributed to the RMSE in the estimation of category parameters. The constant bias in the estimates implies that the estimators may not be consistent, a disturbing finding. Further studies are needed to investigate the effect of bias in the estimates of parameters in polytomous IRT models on the estimation of ability, in the development of item banks, and on adaptive testing.

6.2 Summary and Conclusions for Study II

In Study II, Bayesian estimation was investigated. In particular, the effect of prior distributions on the accuracy of estimation was examined. Prior distributions had an effect on the accuracy of estimates of item parameters in small samples. As can be expected the effect of prior distribution was minimal in large samples. The default priors for slope parameters used in the PARSCALE program resulted in smaller RMSE than that obtained with MMLE in small samples, but yielded larger RMSE of estimates than MMLE in large samples. However, the default priors resulted in smaller RMSE for estimates of the step difficulty parameter than did MMLE. Empirical priors resulted in the largest RMSE for estimates of slope parameters, but produced smaller RMSE than MMLE for estimates of step difficulty parameters.

Bayesian procedures, including empirical priors, yielded smaller variances than MMLE under all conditions. As can be expected, Bayes procedures produced more bias

than MMLE in the estimates of item parameters. Despite the fact that there was more bias with Bayes procedure, it produced smaller RMSE than MMLE. This apparently contradictory finding is the result of the fact that MSE is made up of two parts- sampling variance and bias. Bayes procedure resulted in smaller sampling variance than MML procedure; however, the bias in the Bayes estimates were larger than that found with MMLE. The variance and bias terms combined in such a way as to result in smaller RMSE for Bayes estimates.

In general, the results of study II showed that Bayes procedures provided more accurate estimates of slope parameters with small data sets. However, in order to apply a Bayes procedure prior distributions need to be specified. To investigate if prior distributions based on the data could be useful, the effectiveness of data-based priors was investigated. The transformed proportion of examinees falling in each category was taken as the mean of the distribution for the difficulty parameters, and transformed item-total polyserial correlation was used as the mean of the distribution for the slope parameter.

Empirical, or data-based priors behaved poorly for estimates of the slope parameters. This result may be due to the fact that polyserial correlations are poorly determined in small samples and are poor indicators of slope parameters. Priors on the slope parameters, while having only a modest effect on the accuracy of estimation of slope parameters, had a very positive effect on the accuracy of estimation of the step difficulty parameters. Even the generally poor empirical priors on the slope parameters, produced more accurate estimates of the step difficulty parameters than MMLE.

From the results of item level analysis, it was clear that when the priors matched the true parameter values, very accurate estimates were obtained. In specifying default priors, priors for some items would match the true values. For example, default priors that matched the true values of high-valued slope parameters produced smaller RMSE for the high and medium value slope parameters than MMLE, but produced larger RMSE than MMLE for items with low-valued slope parameters. An interesting and important finding is that *any* prior for slope parameters reduced RMSE of estimates of step difficulty parameters.

The type of item had an effect on the accuracy of estimation. As expected, step difficulty parameters with high or low values were estimated less accurately than those with medium values. This result is probably due to the smaller number of examinees in the extreme categories. Further studies are needed to determine the minimum number of responses in each category to obtain reasonably accurate estimates of the category parameters in polytomous IRT models. Slope parameters with low values were estimated more accurately than those of with high values. However, items with low slope values had a negative effect on the estimation step difficulty parameters especially in small samples.

6.3 Significance of Study

Polytomous IRT models are increasingly used in many situations and accurate estimates of item parameters in polytomous IRT models are critical in practical applications. There are, however, only a few studies have been carried out about the estimation of parameters in polytomous IRT models. The results of this study have

provided valuable information about the properties of various estimators of parameters in the GPCM and the computer program PARSCALE. In particular, results pertaining to the effect of such factors as sample size, test length, the number of categories in each category on the estimates of item parameters will be useful to practitioners who are interested in using the GPCM in assessments about the methods of estimations and conditions under which the GPCM can be successfully applied. The effectiveness of Bayesian procedures in small samples and short tests will be of special importance for performance-based assessment.

6.4 Delimitations and Directions for Further Research

While the present investigation yielded potentially useful findings for practitioners, it also had certain limitations. First of all, this study used the computer program PARSCALE to obtain estimators of parameters in the GPCM. Even though the result of study showed that PARSCALE performed well under various conditions, there was considerable bias in the estimates of the step difficulty parameters under all conditions. To determine the source of bias in the estimates of step difficulty parameters, other computer programs with must be investigated.

Secondly, study II focused on the effect of priors on the slope parameters. Bayes procedures with priors on the slope parameters worked well, except for data-based priors; however, even these priors on the slopes were beneficial for the estimation of step difficulty parameters. While the results obtained in this study II showed that Bayes procedures have the potential for improving the estimation of item parameters in the generalized partial credit model, further research is needed, particularly with respect to

specifying priors for the step difficulty parameters. A hierarchical Bayes procedure as indicated by Swaminathan and Gifford (1982, 1985, and 1986) for estimating parameters may prove to be more useful in the context of the partial credit models. This approach needs to be investigated.

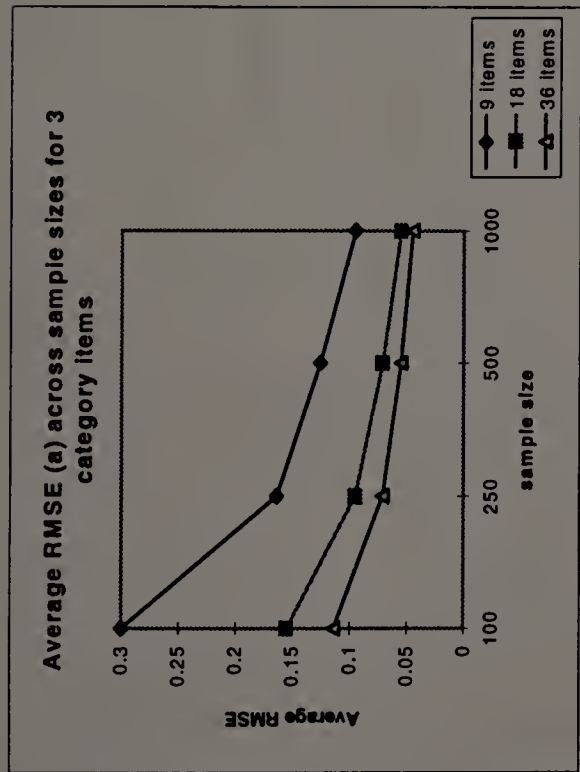
Further study is needed to determine how small a sample size is needed in a response category to obtain reasonable estimates of the category parameters in polytomous IRT models. De Ayala (1995) found that item parameters from the data set with the greatest dispersion of responses across item categories were estimated more accurately than from the data set with the least variability across item categories. The estimation of parameters for categories with few observations tends not to be as accurate as that for categories with relatively more observations. Inaccuracy of parameter estimates may be related to the insufficient number of examinees in response categories and not directly be related to total sample size or ability distribution.

A simulation study with possible combinations of item parameter values is needed to provide more general information for varied conditions. Since this study used the estimates of item parameters from real data set as true item parameter values, those values did not cover all possible combinations of item types.

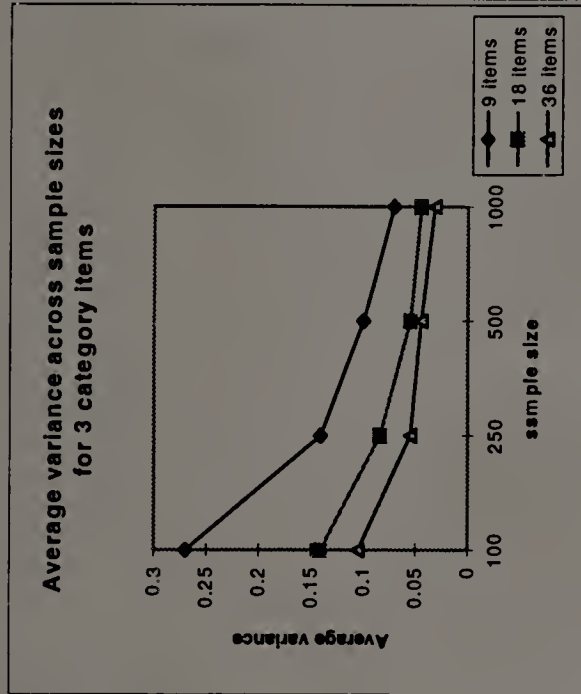
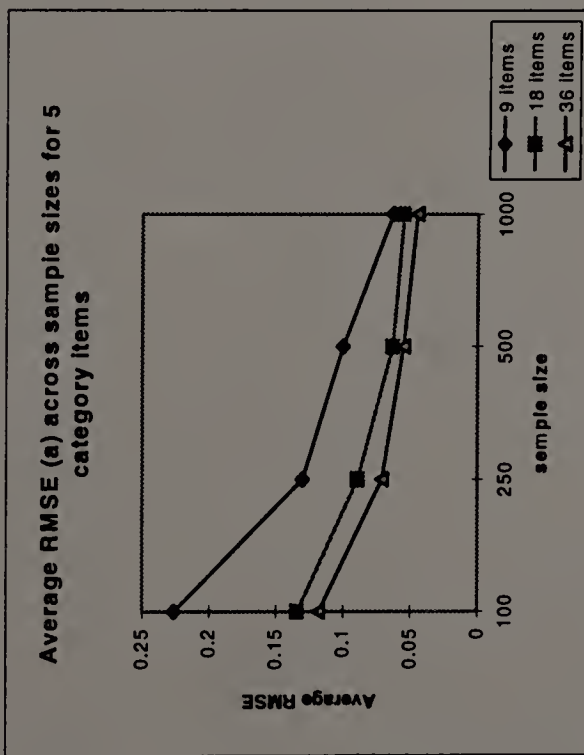
While this study focused on estimates of item parameters because accurate item parameter estimates are critical for such applications as item banking, equating, and studies of differential item functioning, the ultimate purpose of testing is to estimate an examinee's "ability" or proficiency level. It is necessary, therefore, to investigate the conditions under which accurate estimates of ability parameters is obtained in polytomous IRT models. In order to estimate ability parameters, it has to be assumed that

accurate values of item parameters are available. The current study has shown that item parameters are estimated with error. The effect of item parameter estimate error on ability parameter estimates is not known and a detailed investigation of the effects of errors in item parameter estimation on ability estimation needs to be undertaken.

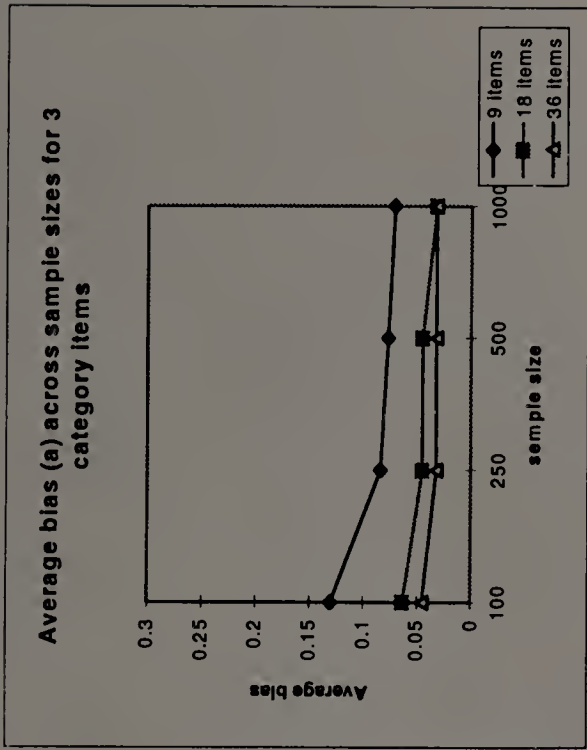
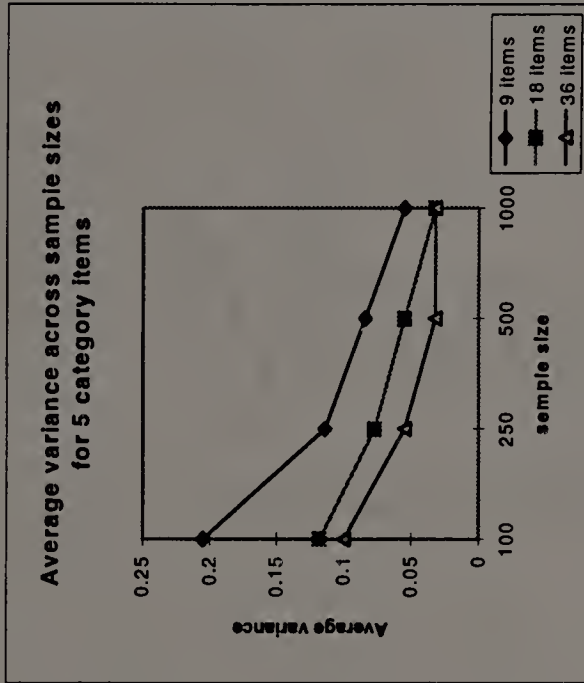
APPENDIX
ADDITIONAL FIGURES



(a)



(b)



(c)

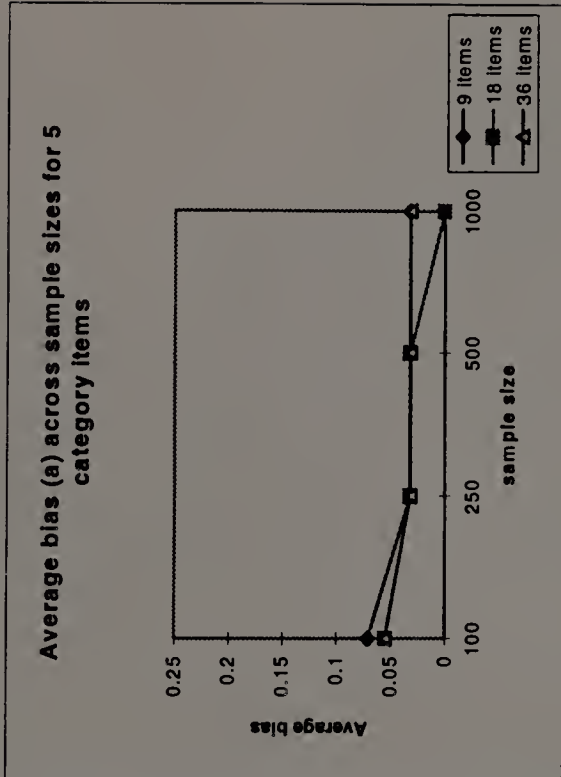


Figure A.1. Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on uniform distribution

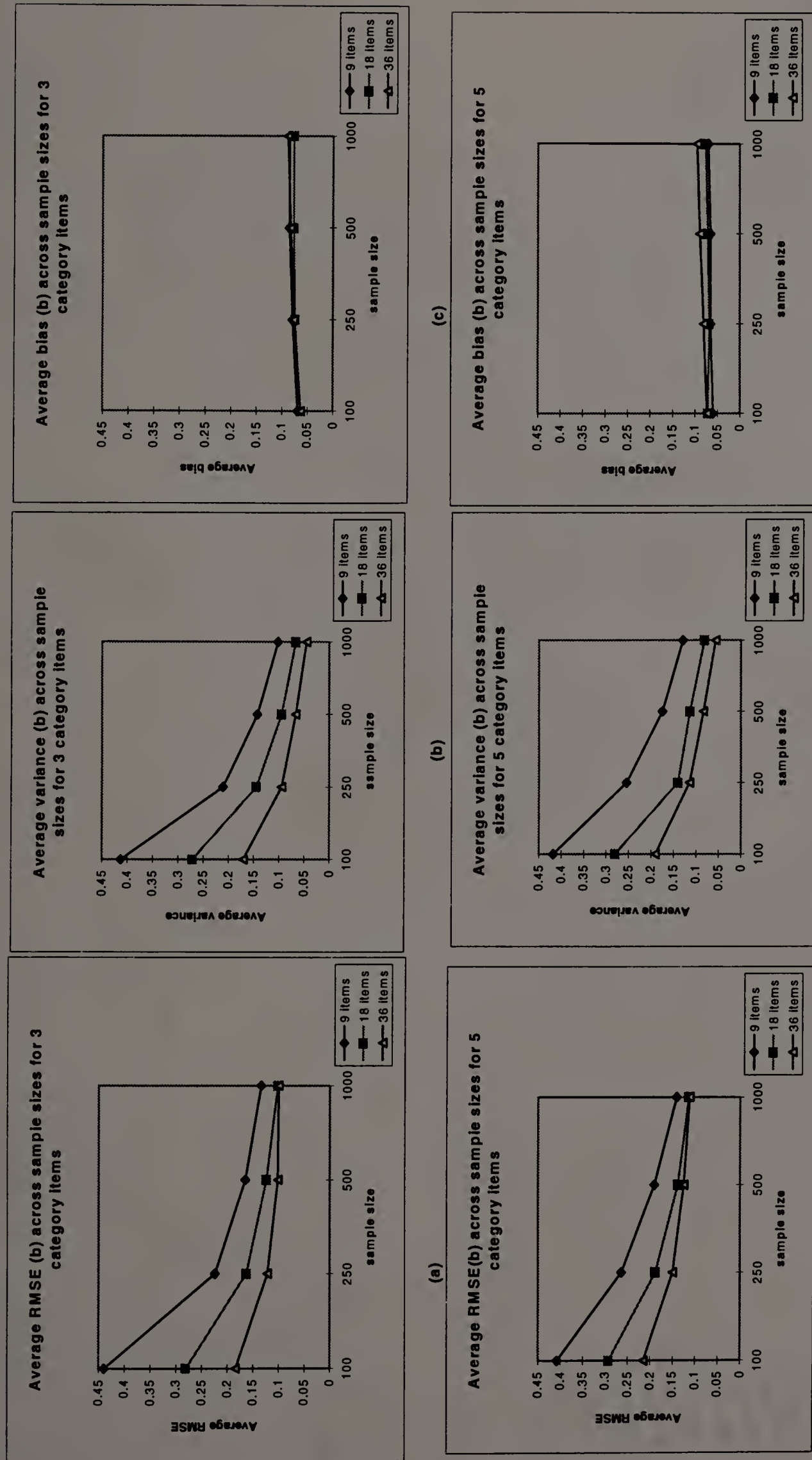
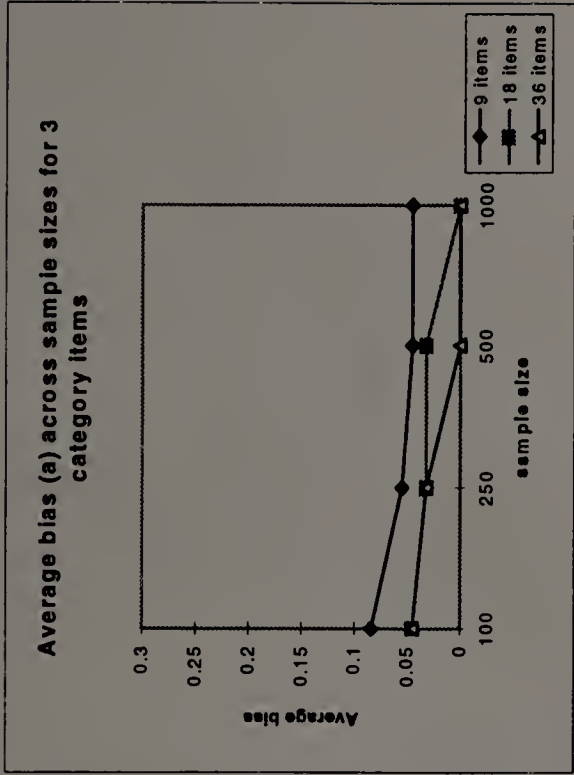
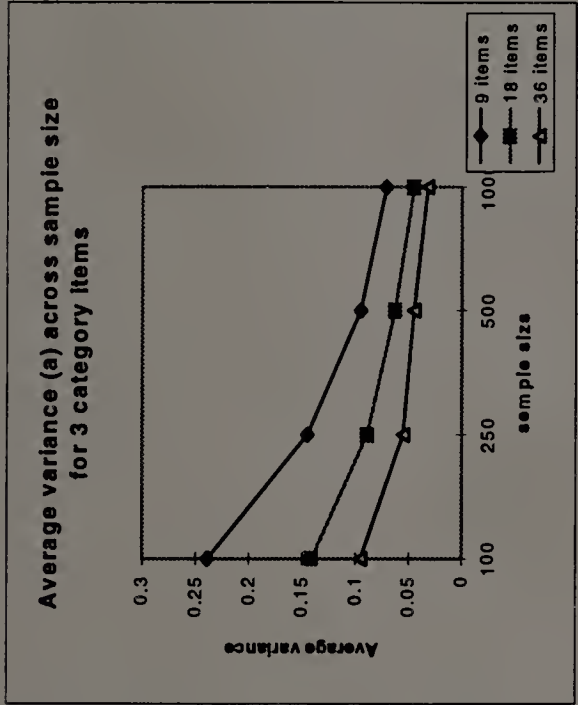
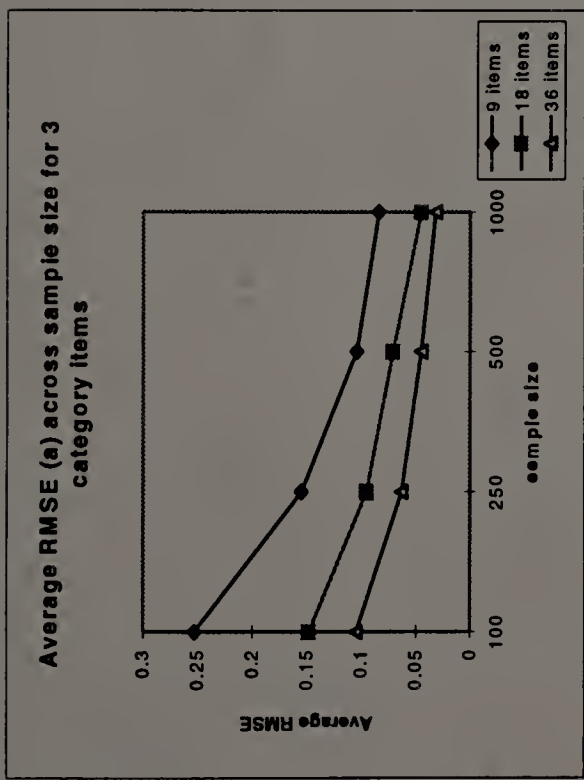
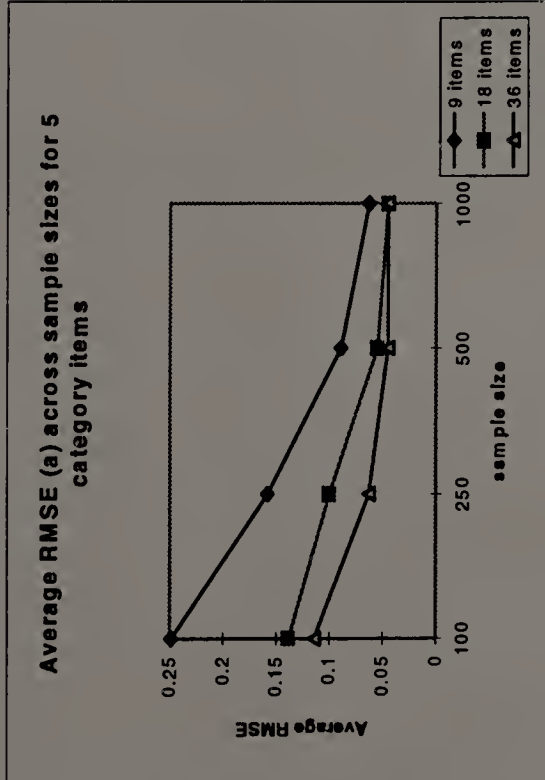


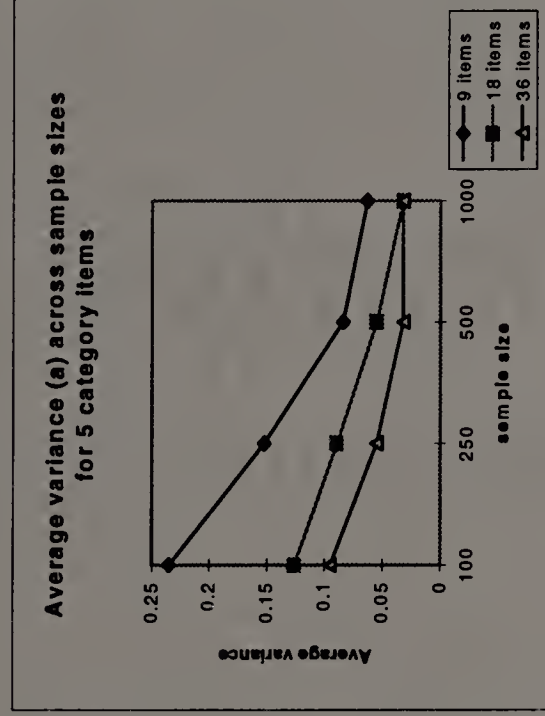
Figure A.2. Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on uniform distribution



(a)



(b)



(c)

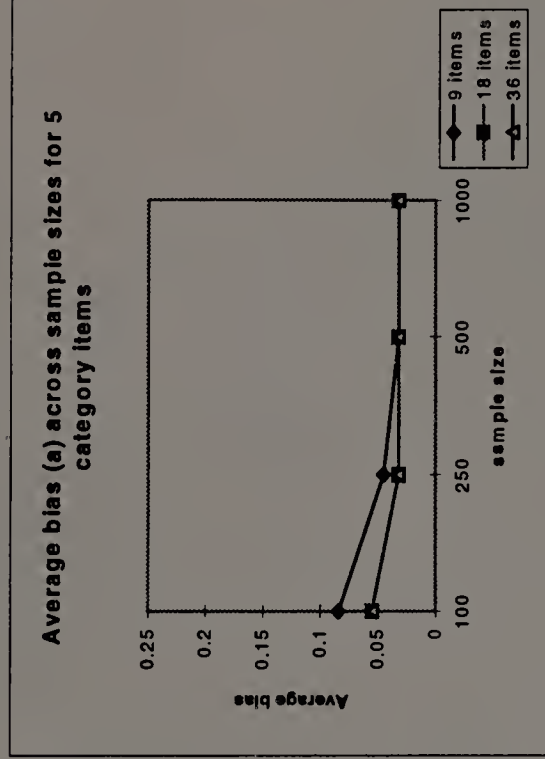


Figure A.3. Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on positively skewed distribution

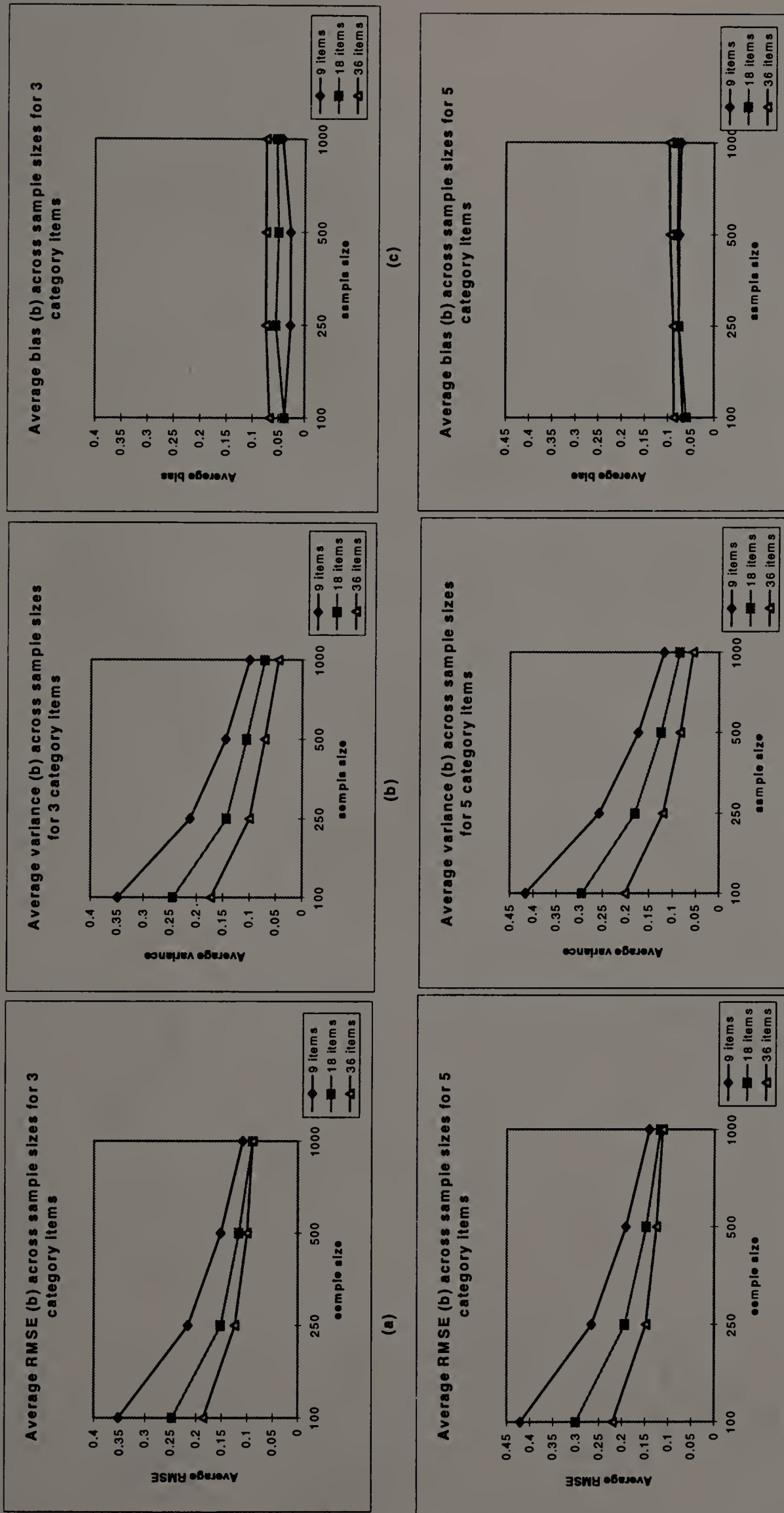
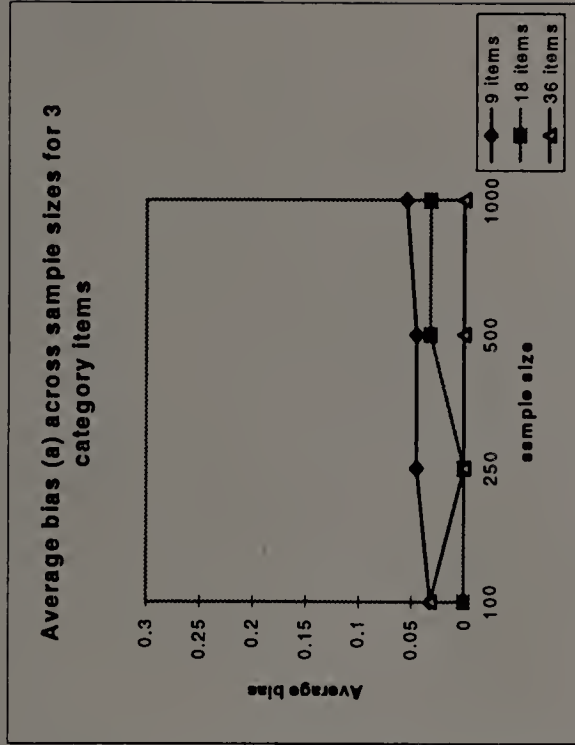
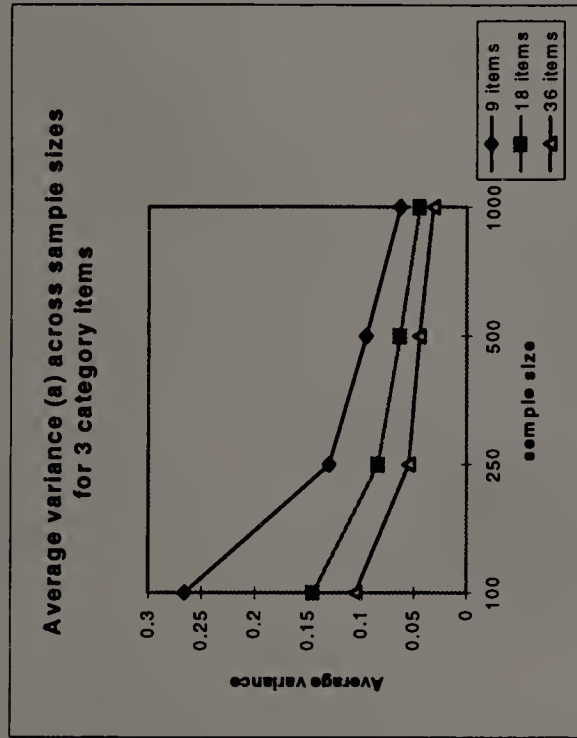
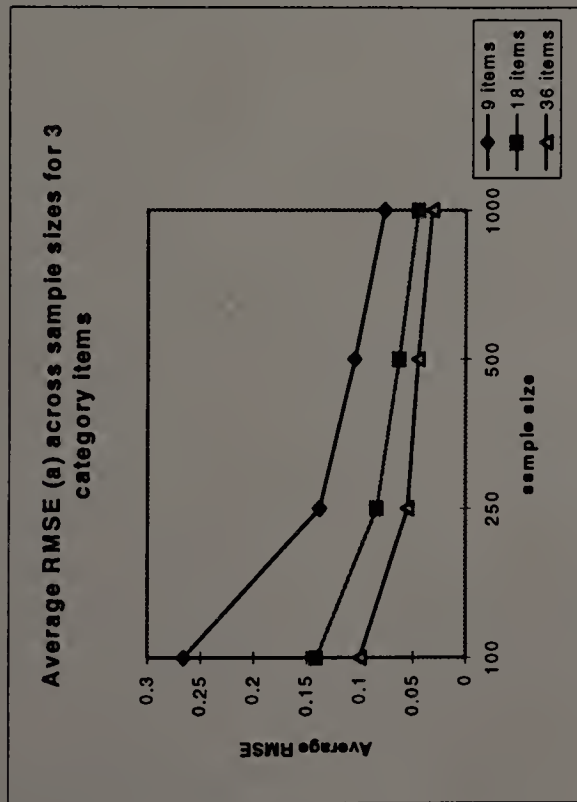
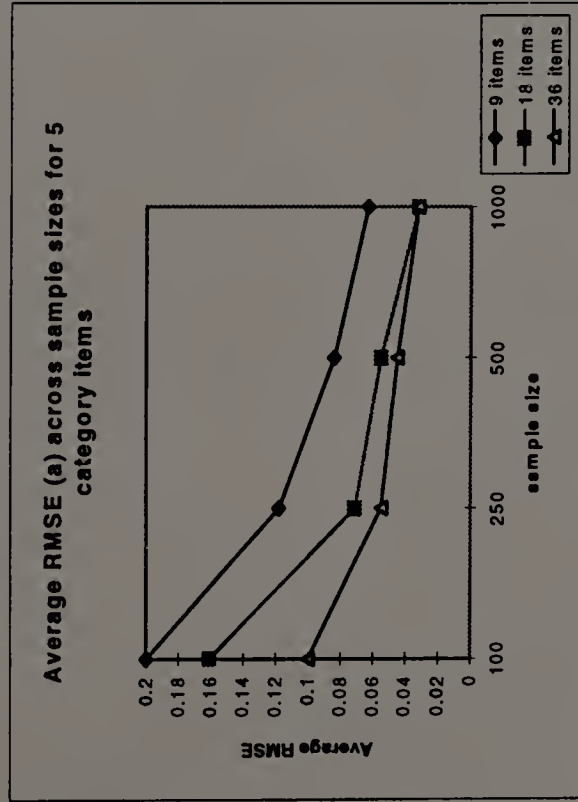


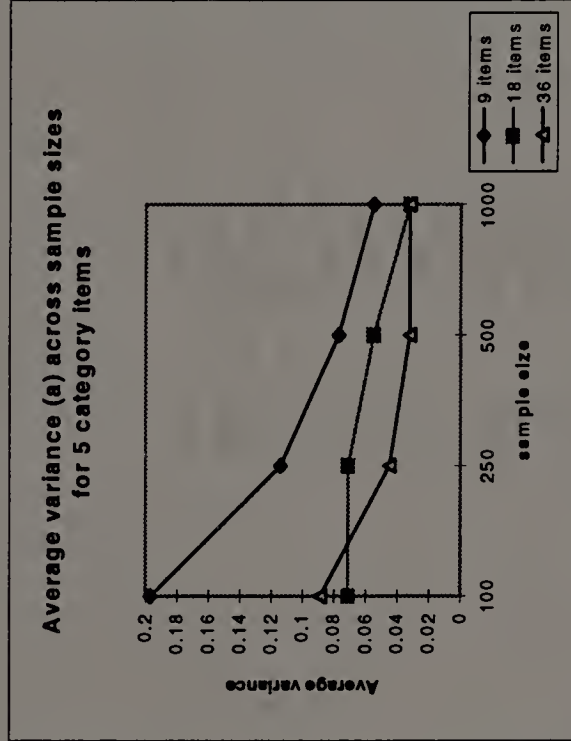
Figure A.4. Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on positively skewed distribution



(a)



(b)



(c)

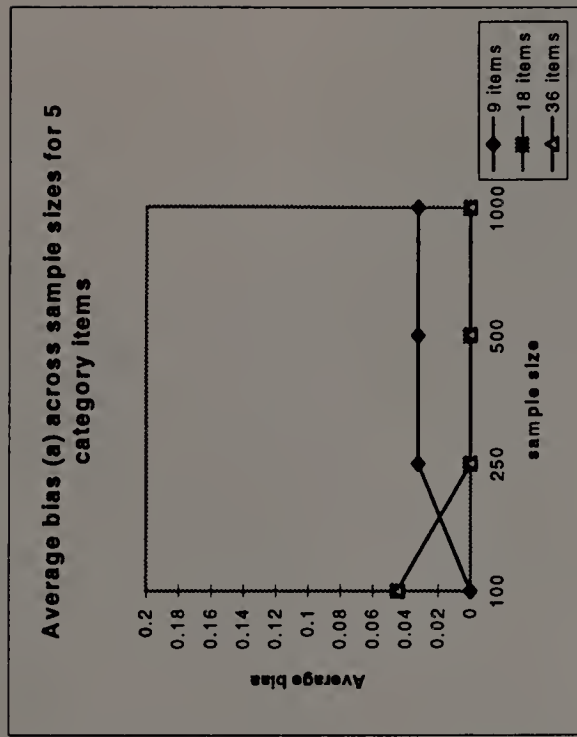
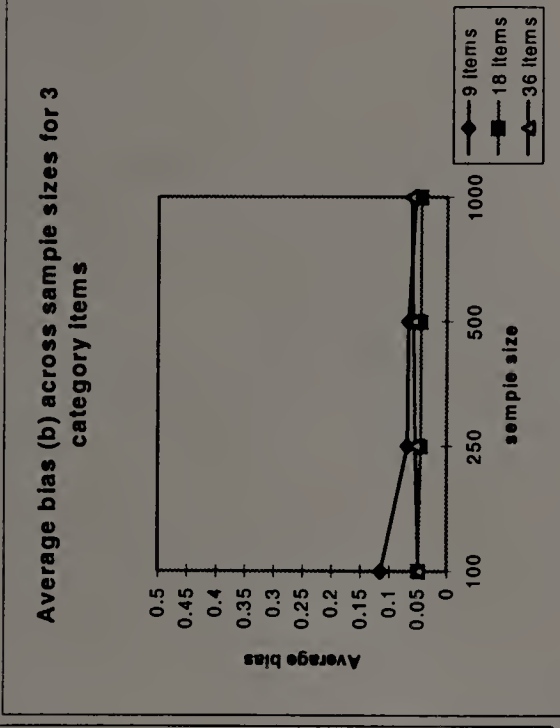
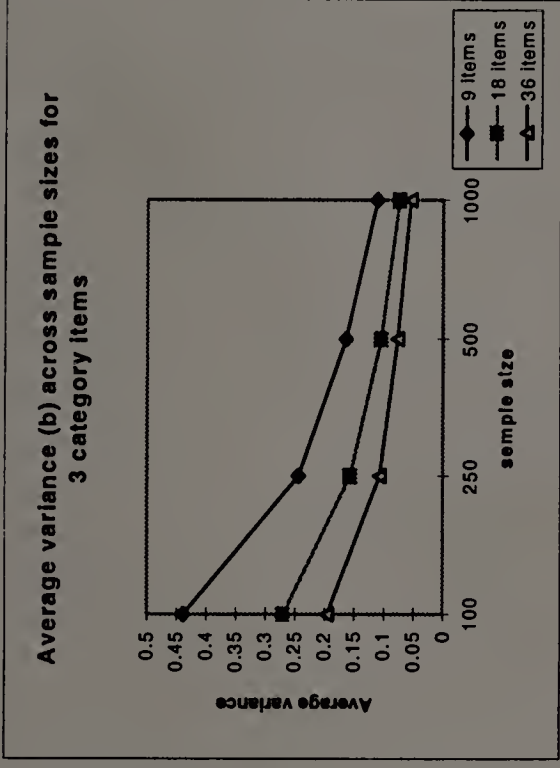
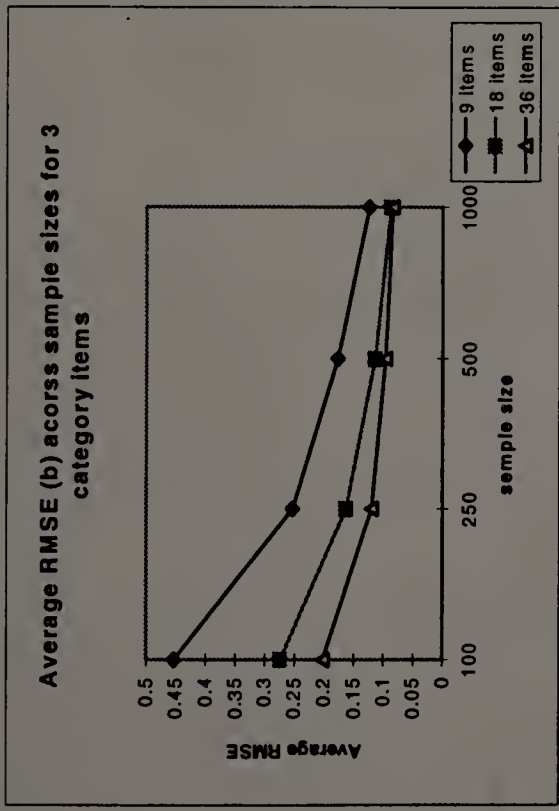
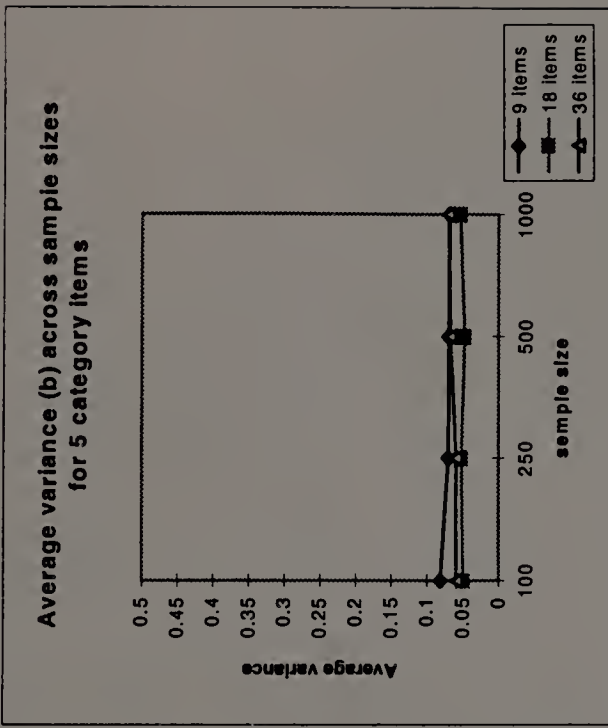
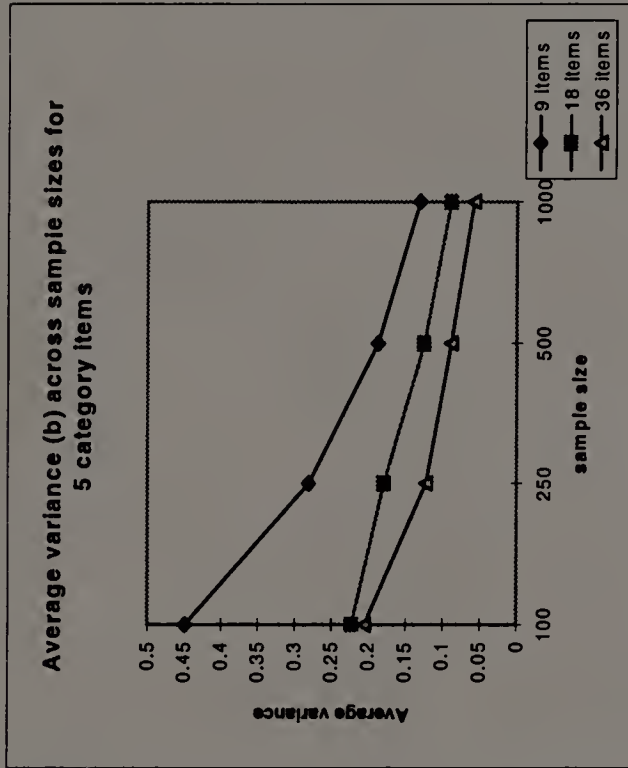
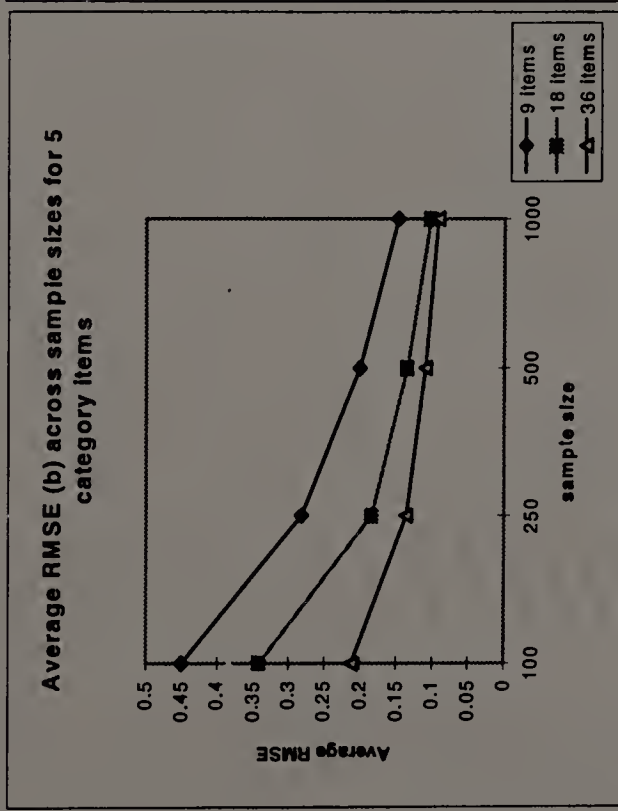


Figure A.5. Average RMSE, variance, and bias of estimates of slope parameters across sample sizes and test lengths for 3 and 5 category items based on negatively skewed distribution



(a)



(b)

(c)

Figure A.6. Average RMSE, variance, and bias of estimates of step difficulty parameters across sample sizes and test lengths for 3 and 5 category items based on negatively skewed distribution

REFERENCES

- Andersen, E. B. (1973). Conditional inference for multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 26, 42-54.
- Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 42, 69-81.
- Baker, F. B. (1992). Item response theory : Parameter estimation techniques. New York: Marcel Dekker, Inc.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addition-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. Psychometrika, 46, 443-459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.
- Brown, R. L.(1991). The effect of collapsing ordered polytomous scales on parameter estimates in structural equation measurement models. Educational and Psychological Measurement, 51, 317-328.
- Carlson, J. E. (1996, April). Information provided by polytomous and dichotomous items on certain NAEP instruments. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Choi, S. W., Cook, K. F., & Dodd, B. G. (1996, April). Parameter recovery for the partial credit model using MULTILOG. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1996.
- De Ayala, R. J. (1995, April). Item parameter recovery for the nominal response model. Paper presented at the annul meeting of the American Education Research Association, San Francisco, CA.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. The journal of the Royal Statistical Society, Series B, 39, 1-38.

- Dodd, B. G., De Ayala, R. J., & Koch, R. W. (1995). Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 19, 5-22.
- Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the GPCM. Journal of Educational Measurement, 31, 295-311.
- Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.
- Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. Applied Psychological Measurement, 13, 77-90.
- Embretson, S. (1984). A general latent trait model for response processes. Psychometrika, 49, 175-186.
- Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. Applied Psychological Measurement, 11, 33-44.
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and application. Boston: Kluwer-Nijhoff.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA : SAGE Publications, Inc.
- Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement, 15, 279-291.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, 6, 249-260.
- Lecointe, D. A. (1995, April). How the collapsing of categories impact the item information function in polytomous item response theory. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Lim, R. G., & Drasgow F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. Journal of Applied Psychology, 75, 164-174.

- Lindley, D. V., & Smith, A. F. (1972). Bayesian estimates for the linear model. Journal of the Royal Statistical Society, Series B, 34, 1-41.
- Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, N0. 7.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 23, 157-162.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.
- McKinley, R. L., Reckase, M. D. (1982). The Use of General Rasch Model with Multidimensional Item Response Data. Iowa City, IA: American College Testing.
- Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software, Inc.
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.
- Muraki, E., & Bock, R. D. (1993). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago, IL: Scientific Software.
- Park, C. , & Swaminathan, H. (1996). POLYGEN : A Fortran program for generating polytomous IRT models. Amherst, MA: School of Education, University of Massachusetts.
- Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomous scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19, 23-38.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34 (4, Whole No. 17)

- Samejima, F. (1972). A general model for free-response data. Psychometrika Monograph No. 18.
- Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14, 299-311.
- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. Applied Psychological Measurement, 16, 1-16.
- Swaminathan, H. (1983). Parameter estimation in item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 24-44). Vancouver, BC: Educational Research Institute of British Columbia.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-192.
- Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50, 349-364.
- Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrika, 51, 589-601.
- Sympson, J. B. (1983). A new IRT model for calibrating multiple choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.
- Thissen, D. (1976). Information in wrong responses to the Raven progressive matrices. Journal of Educational Measurement, 13, 201-214.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47, 175-186.
- Thissen, D. (1991). MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory. Mooresville, IN: Scientific Software.
- Thissen, D., & Steinberg, L. (1984). A model for multiple choice items. Psychometrika, 49, 501-519.
- Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. Applied Psychological Measurement, 12, 53-67.

Walker-Bartnick, L. A. (1990). An investigation of factors affecting invariance of item parameter estimates for the partial credit model. Ann Arbor, Michigan: UMI Dissertations and Theses Information Services.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

